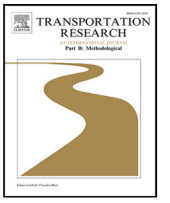
Transportation Research Part B 192 (2025) 103129

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb



Revisiting McFadden's correction factor for sampling of alternatives in multinomial logit and mixed multinomial logit models

Thijs Dekker [a,*], Prateek Bansal [b], Jinghai Huo [c]

[a] Institute for Transport Studies, University of Leeds, UK
[b] Department of Civil and Environmental Engineering, National University of Singapore, Singapore
[c] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85281, USA

ARTICLE INFO

Keywords:
Multinomial logit
Mixed multinomial logit
Sampling of alternatives
Information loss
Bayesian estimation

ABSTRACT

When estimating multinomial logit (MNL) models where choices are made from a large set of available alternatives computational benefits can be achieved by estimating a quasi-likelihood function based on a sampled subset of alternatives in combination with '*McFadden's correction factor*'. In this paper, we theoretically prove that McFadden's correction factor minimises the expected information loss in the parameters of interest and thereby has convenient finite (and large sample) properties. That is, in the context of Bayesian estimation the use of sampling of alternatives in combination with McFadden's correction factor provides the best approximation of the posterior distribution for the parameters of interest irrespective of sample size. As sample sizes become sufficiently large consistent point estimates for MNL can be obtained as per McFadden's original proof. McFadden's correction factor can therefore effectively be applied in the context of Bayesian MNL models. We extend these results to the context of mixed multinomial logit models (MMNL) by using the property of data augmentation in Bayesian estimation. McFadden's correction factor minimises the expected information loss with respect to the augmented individual-level parameters, and in turn also for the population parameters characterising the shape and location of the mixing density in MMNL. Again, the results apply to finite and large samples and most importantly circumvent the need for additional correction factors previously identified for estimating MMNL models using maximum simulated likelihood. Monte Carlo simulations validate this result for sampling of alternatives in Bayesian MMNL models.

1. Introduction

Recent works in transportation, environmental economics and marketing (Daly et al., 2014; Guevara and Ben-Akiva, 2013a,b; Keane and Wasi, 2016; Sinha et al., 2018; Tsoleridis et al., 2022; Von Haefen and Domanski, 2018) have renewed interest in the sampling of alternatives in the context of discrete choice modelling. Sampling of alternatives reduces the computational challenge of evaluating the denominator of the logit choice probability for large choice sets by only making use of a smaller subset of sampled alternatives including the chosen alternative. The benefit of sampling alternatives stems from a significant reduction in the estimation time of large-scale choice models. Evaluating the logit formula over a subset of alternatives by default overestimates the choice probability which in turn may have unintended consequences for estimating model parameters. McFadden (1978) already proved for multinomial logit (MNL) models that an application of a sampling correction to the utility function of the sampled alternatives results in consistent parameter estimates.

* Corresponding author.
E-mail addresses: t.dekker@leeds.ac.uk (T. Dekker), prateekb@nus.edu.sg (P. Bansal), jinghaih@asu.edu (J. Huo).

https://doi.org/10.1016/j.trb.2024.103129
Received 28 September 2023; Received in revised form 24 September 2024; Accepted 10 November 2024

Available online 4 December 2024

Most discrete choice applications nowadays apply model specifications beyond MNL. Models from the Multivariate Extreme Value (MEV) family, such as the nested logit model (Daly, 1987), and mixed multinomial logit (MMNL) models (Revelt and Train, 1998) have become state of the art. Guevara and Ben-Akiva (2013b) and Guevara and Ben-Akiva (2013a) proved that McFadden (1978)'s proposed correction factor forms part of the solution to implement sampling of alternatives within these more advanced models. Consistent parameter estimates can be obtained when in addition to McFadden's correction term, the analyst also corrects for the imperfect representation of the LogSum in MEV in the sampled model, and the latent nature of individual-level parameters in MMNL.

Guevara and Ben-Akiva (2013a) highlight in their conclusions the need to study sampling of alternatives in the context of finite sample sizes and alternative estimation strategies. Von Haefen and Domanski (2018) argue consistency of parameter estimates and welfare measures of latent class models using the Expectation–Maximisation (EM) algorithm in combination with sampling of alternatives. They also empirically show good performance for relatively small sizes of sampled choice sets, up to 5% of the full choice set size, but do not address the issue of finite sample sizes. In this paper, we study sampling of alternatives in the context of Bayesian estimation routines for MNL and MMNL models. Since the Bayesian approach does not rely on asymptotically large sample sizes, our paper addresses both aspects highlighted by Guevara and Ben-Akiva (2013a) simultaneously.

With the advancements in computational resources and approximate inference, Bayesian estimation has gained popularity in the discrete choice modelling literature (Bansal et al., 2020). Bayesian estimation is particularly fruitful when latent constructs are included in the likelihood function. Significant reductions in estimation time over classical maximum simulated likelihood (MSL) approaches are typically obtained due to augmenting latent constructs (Tanner and Wong, 1987; Train, 2009), such as individual-level parameters in mixed logit models or class membership in latent class models. The joint implementation of sampling of alternatives and Bayesian estimation may therefore lead to additional time savings and spur the implementation of more flexible discrete choice models in travel demand models characterised by large choice sets.

In this paper, we take a somewhat unconventional approach by revisiting three key papers in the sampling of alternatives literature, respectively McFadden (1978), Keane and Wasi (2016) and Guevara and Ben-Akiva (2013a). Our theoretical analysis uses key insights from these three papers and builds upon them to arrive at new insights in the context of finite sample sizes, Bayesian estimation methods, and MMNL models. Section 2 discusses McFadden (1978)'s correction factor in the context of MNL. We provide an intuitive explanation of the role of the correction factor in the corrected MNL choice probability under the sampling of alternatives. We decompose McFadden (1978)'s correction factor into (i) a correction for overestimating the MNL choice probability due to using a smaller subset of alternatives, and (ii) a correction for which subset of alternatives is contrasted through utility differences. Only the latter determines the extent to which we learn about the parameters of interest in MNL. Building on the work of Keane and Wasi (2016), we show that these two components comprise the loss of information (or information divergence) between the 'true' and 'sampled' MNL log-likelihood. We show that only the second component – operating through the denominator of the MNL choice probability – is relevant for estimation purposes. According to our results, McFadden (1978)'s correction factor not only results in consistent parameter estimates but also minimises the expected loss of information with respect to the parameters of interest in MNL. This result generalises and puts context to Keane and Wasi (2016)'s conclusions on the expected information divergence under sampling of alternatives in MNL.

The outcome that McFadden (1978)'s correction factor minimises the expected loss of information with respect to the parameters of interest in MNL is central to implementing sampling of alternatives in Bayesian MNL and MMNL models (and thereby establishing the desirable finite sample properties of McFadden (1978)'s correction factor) in Sections 3 and 4, respectively. In these two sections, we evaluate the performance of McFadden (1978)'s correction factor across the entire posterior distribution irrespective of sample size, and in relation to the consistency of Bayesian point estimates for large sample sizes. We position our results against the work of Guevara and Ben-Akiva (2013a). Section 5 supports our theoretical analysis with the use of Monte Carlo simulations. Section 6 concludes and sets out a pathway for future research.

2. Sampling of alternatives in classical MNL models

2.1. Sampling of alternatives and McFadden's correction factor

This paper follows the conventional micro-econometric approach to individual decision-making. Individual n is assumed to select alternative i from the choice set C_n when it generates the highest level of indirect utility

$$U_{in} = V_{in} + \epsilon_{in} = V(X_{in}; \beta) + \epsilon_{in}. \tag{1}$$

Indirect utility U_{in} in Eq. (1) comprises a deterministic part V_{in} and an additive unobserved stochastic part ϵ_{in}. The deterministic part is a function of explanatory variables X_{in} and associated parameters β, and typically takes a linear form. The stochastic term ϵ_{in} is assumed to be independently and identically distributed across alternatives and individuals.

Assume that the 'true' data generating process takes the form of the logit model such that ϵ_{in} follows a Type 1 Extreme Value distribution. Accordingly,

$$P(i|C_n; \beta) = \frac{\exp(V_{in})}{\sum_{j \in C_n} \exp(V_{jn})} = \frac{1}{\sum_{j \in C_n} \exp(V_{jn} - V_{in})} \tag{2}$$

describes the probability that individual n will select alternative i.[1] The denominator highlights the computational challenge of applying MNL to large choice sets.

Sampling of alternatives reduces the computational burden by specifying a quasi-likelihood function approximating Eq. (2) using the smaller choice set D_n. D_n is a subset of C_n and includes a number of randomly sampled alternatives, besides the chosen alternative i. Let

$$P(i|D_n;\beta)^{\dagger} = \frac{\exp(V_{in})}{\sum_{j\in D_n}\exp(V_{jn})} = \frac{1}{1+\sum_{j\neq i\in D_n}\exp(V_{jn}-V_{in})} \tag{3}$$

define the *uncorrected* sampled choice probability. The first aspect revealed by Eq. (3) is that by using the sampled choice set D_n, which is a subset of C_n, a smaller number of alternatives will be included in the denominator and by default $P(i|D_n;\beta)^{\dagger}$ *overestimates* the true choice probability $P(i|C_n;\beta)$. Note that implicitly fewer utility differences are evaluated. These utility differences provide information on the parameters of interest. Accordingly, $P(i|D_n;\beta)^{\dagger}$ may be associated with bias in its corresponding parameter estimates $\hat{\beta}$ because the sampling protocol determines the likelihood for each alternative to enter the subset D_n, and thereby the utility differences evaluated to learn about β.

McFadden (1978) proved consistent parameter estimates are obtained by adding the correction factor $ln(\pi(D_n|j))$ to the indirect utility function of each alternative $j \in D_n$, where $\pi(D_n|j)$ denotes the probability of sampling the choice set D_n conditional on alternative j being the chosen alternative. The only requirement for consistency is a trivial condition on this sampling probability — which McFadden (1978) labels as *Positive Conditioning*, i.e. $\pi(D_n|j) > 0 \;\forall\; j \in D_n$. A positive conditional probability of sampling the choice set D_n allows evaluating the correction factor $ln(\pi(D_n|j))$. Appendix A replicates McFadden (1978)'s original proof.

Define McFadden (1978)'s corrected MNL choice probability by

$$P(i|D_n;\beta) = \frac{exp\left(V_{in}+ln(\pi(D_n|i))\right)}{\sum_{j\in D_n} exp\left(V_{jn}+ln(\pi(D_n|j))\right)} = \frac{1}{1+\sum_{j\neq i\in D_n} exp\left(V_{jn}-V_{in}+ln(\pi(D_n|j))-ln(\pi(D_n|i))\right)}. \tag{4}$$

In Eq. (4) $\pi(D_n|i)$ acts as a constant in the denominator scaling down the choice probability and correcting for the overestimation of the choice probability. It is commonly assumed that the sampling protocol $\pi(D_n|i)$ is independent of β and accordingly the term only rescales the likelihood (i.e. add a negative constant to the log-likelihood). It does not influence where the likelihood function is maximised. This term therefore does not correct for any potential bias in the parameters of interest. Therefore, $\pi(D_n|j)$ corrects for the potential bias induced by the sampling protocol as to which utility differences are contrasted under sampling of alternatives.

2.2. Sampling protocols

Besides including the chosen alternative, the analyst needs to make several practical considerations when sampling the remainder of D_n. The sampling protocol determines the (expected) size of the sampled choice set and thereby introduces a trade-off between speed and accuracy. With smaller D_n, the estimation will be computationally more attractive, but the approximation of the denominator in the choice probability is likely to become less accurate. That is, the number of utility differences evaluated – which provide the necessary information to estimate the parameters of interest – is reducing linearly with the size D_n. The balance between computational costs and sampling error remains an empirical matter. Nerella and Bhat (2004) suggest that an eighth of the size of the full choice set should be used as minimum, but that a fourth is desirable.

The interested reader is referred to Ben-Akiva and Lerman (1985) for a description of various sampling protocols to obtain the target size of D_n, including the corresponding correction factors. Monte Carlo simulations conducted by Daly et al. (2014) highlight that independent and with-replacement sampling allows for efficient estimation and easy calculation of $\pi(D_n|j)\;\forall j \in D_n$. These two sampling protocols, however, result in variation in the size of D_n across observations, whereas the size of D_n is fixed under the sampling without replacement. The calculation of the correction factor $\pi(D_n|j)$ is, typically, more complex under the sampling without replacement and estimation is less efficient. Stratified sampling, where a fixed number of alternatives is sampled from a number of strata, suffers from the same challenges as sampling without replacement. Whereas Daly et al. (2014) discard stratified sampling due to poor estimation performance, Guevara and Ben-Akiva (2013b) apply a stratified importance sampling protocol without replacement in their Monte Carlo simulation on nested logit models in order to ensure that sufficient alternatives are sampled per nest.

The next consideration is assigning a sampling probability to each *alternative* in the choice set. It is at the level of alternatives that standard uniform or importance sampling probabilities can be assigned. In the former, sampling probabilities are constant across alternatives whereas in the latter they vary based on some specified rule. For example, in destination choice, alternatives close to the chosen alternative may receive a higher sampling probability as they are more likely to be chosen. Typically, defining the non-constant sampling probabilities at the level of alternatives is done by relying on results from previous studies such that no endogeneity is introduced. For McFadden (1978)'s correction factor, the resulting conditional sampling probabilities of the *set* $\pi(D_n|j)\;\forall j \in D_n$ are relevant. Different combinations of sampling probabilities and protocols at the level of the alternatives may result in the same conditional sampling probabilities at the level of the set D_n.

[1] To improve the clarity of notation, we remove the conditioning on X_n from all probability statements. X_n typically represents a set of exogenous variables not associated with any form of stochasticity. If any such stochasticity is present, for example in a latent variable model, our Bayesian results still apply as long as the stochasticity is independent of β.

An attractive choice for the conditional sampling probability at the level of the set is uniform conditioning whereby an equal conditional probability of being sampled is assigned to each eligible D_n such that $\pi(D_n|j) = \pi(D_n|k)\ \forall j,k$. Uniform conditioning adds a constant to the indirect utility of each alternative in Eq. (4) and thereby the correction factor cancels out in the corrected choice probability. Eq. (3) thus provides consistent estimates under uniform conditioning without the need to adjust the indirect utility function. It is intuitive that no correction factor is needed under uniform conditioning to avoid bias because the sampling protocol does not steer the analysis towards a specific set of alternatives from the full choice set. Uniform conditioning does not make it more or less likely that certain utility differences are evaluated. As the sample size increases, the randomness of uniform conditioning increases the probability that all relevant utility differences are studied in a balanced way across the sample which is desirable to obtain consistent parameter estimates.

An undesirable feature of uniform conditioning, however, is that in decision problems with a really large number of alternatives (e.g. destination choice), the inclusion of irrelevant options in the sampled choice set may require a prohibitively large sample size to achieve proper estimation. As noted above, importance sampling overcomes this issue by assigning a higher sampling probability to alternatives that are a priori more likely to be chosen. Importance sampling at the level of the alternative often results in a non-uniform conditional sampling probability such that $\pi(D_n|j) \neq \pi(D_n|k)\ \forall j,k$. In this context, McFadden (1978)'s correction factor is, however, needed as per Eq. (4) to correct for the potential bias induced by steering the analysis towards a specific set of alternatives from the full choice set. In order to successfully evaluate $ln(\pi(D_n|j))\ \forall j \in D_n$ McFadden (1978) has defined the positive conditioning requirement. Uniform conditioning is considered a specific case of the more generic positive conditioning.

2.3. Information divergence

The difference between the quasi log-likelihood based on Eq. (4) and the correct log-likelihood evaluating the full choice set is studied in more detail by Keane and Wasi (2016). Similar to McFadden (1978)'s proof, Keane and Wasi (2016) study the information divergence from an *ex ante* perspective — taking expectations over all possible chosen alternatives and all possible sampled choice sets. Appendix B analytically derives their two main conclusions for the case of uniform conditioning: (i) the expected information divergence is positive, and (ii) the expected information divergence is minimised at the true parameters. In Appendix B, we furthermore show that this result does not generalise to the case of positive conditioning. Instead, we arrive at similar but slightly adjusted conclusions in relation to McFadden (1978)'s correction factor under positive conditioning: (1) the expected loss in information *with respect to the parameters of interest* is positive, and (2) the expected loss in information *with respect to the parameters of interest* is minimised at the true parameters.

The explanation for adjusting the conclusions from Keane and Wasi (2016) can be traced back to the difference in the role of $\pi(D_n|i)$ and $\pi(D_n|j)$ in McFadden's corrected choice probability. With reference to Section 2.1, $\pi(D_n|i)$ only rescales the choice probability independently of β, whereas $\pi(D_n|j)$ controls for potential biases in the parameter estimates due to the selected sampling protocol. For uniform conditioning, the sign of the overall expected information divergence is guaranteed to be *positive* because the size of the denominator is smaller in the sampled model and $\pi(D_n|i)$ and $\pi(D_n|j)$ cancel each other out. Under uniform conditioning, the expected information divergence is therefore equivalent to the expected information loss with respect to the parameters of interest. Under positive conditioning $\pi(D_n|i)$ and $\pi(D_n|j)$ no longer cancel each other out. Since $ln(\pi(D_n|j)) < 0\ \forall j \in D_n$, $\pi(D_n|j)$ further inflates the choice probabilities in the sampled model relative to the true model and increases the *positive* expected information divergence. $\pi(D_n|i)$, however, has the opposite effect on the information divergence because it decreases the corrected choice probability. Although unlikely, an over-correction for the overestimation of the MNL choice probability may in theory occur under positive conditioning causing the sampled likelihood to be lower than the true likelihood. A positive sign of the expected information divergence can therefore not be guaranteed for positive conditioning and Keane and Wasi (2016)'s first result therefore does not transfer to the more generic case of positive conditioning.

Since $\pi(D_n|i)$ is irrelevant for estimation purposes we can, however, ignore this part of the overall expected information divergence. What we are left with is a *positive* expected information loss with respect to the parameters of interest due to contrasting fewer utility differences based on a specific sampling protocol $\pi(D_n|j)$. Appendix B shows that at the true parameters, this expected information loss is minimised under *both* uniform and positive conditioning. Hence, McFadden (1978)'s correction factor not only results in consistent parameter estimates under uniform and positive conditioning in MNL but also minimises the expected loss of information with respect to the parameters of interest.

3. Sampling of alternatives in Bayesian MNL models

When moving from classical to Bayesian estimation, the original objective – recovering the true point parameter estimates β^* – changes to recovering the posterior distribution. The posterior distribution

$$p(\beta|Y,C) = \frac{p(\beta)p(Y|C;\beta)}{p(Y|C)} = \frac{p(\beta)\prod_{n=1}^{N}\frac{\exp(V_{in})}{\sum_{j\in C_n}\exp(V_{jn})}}{\int_{\beta}p(\beta)\prod_{n=1}^{N}\frac{\exp(V_{in})}{\sum_{j\in C_n}\exp(V_{jn})}d\beta} \tag{5}$$

is described as a function of the prior distribution $p(\beta)$, the likelihood $p(Y|C;\beta)$ and the marginal likelihood $p(Y|C)$ - where C denotes the set of full choice sets C_n across all observations $n = 1, 2, \ldots, N$ and Y the vector of all observed choices in the sample.

- The prior distribution $p(\beta)$ summarises all information about the parameters of interest *before* collecting the data. Given our interest in using a quasi-likelihood function to approximate the true likelihood function, containing the *same* β parameters, there is no reason to assume that the prior differs between the 'true' and 'sampled' model.
- The likelihood $p(Y|C;\beta)$ is *identical* to the likelihood function in classical estimation and describes the probability of observing the choices Y in the sample for a given β and full choice set C. The same applies to the sampled likelihood function introduced below.
- The marginal likelihood $p(Y|C)$ represents the expected likelihood of the model and thereby describes how well the model fits the data across all possible values of β.

Since β is integrated out of the marginal likelihood, it does not contain information on the parameters of interest and only scales the posterior distribution to ensure that the density integrates to one. For estimation purposes, it is therefore often considered that the posterior is proportional to the prior and the likelihood, i.e. $p(\beta|Y,C) \propto p(\beta)\cdot p(Y|C;\beta)$. In a similar vein, any terms that are multiplicatively unrelated to β, and thereby only scale the posterior, can be removed since they do not contain information on the parameters of interest. If we now let

$$p(\beta|Y,D) = \frac{p(\beta)\cdot p(Y|D;\beta)}{p(Y|D)} = \frac{p(\beta)\cdot \prod_{n=1}^{N} \frac{\exp(V_{in}+ln(\pi(D_n|i)))}{\sum_{j\in D_n}\exp(V_{jn}+ln(\pi(D_n|j)))}}{\int_\beta p(\beta)\cdot \prod_{n=1}^{N} \frac{\exp(V_{in}+ln(\pi(D_n|i)))}{\sum_{j\in D_n}\exp(V_{jn}+ln(\pi(D_n|j)))} d\beta} \tag{6}$$

describe the posterior density under McFadden (1978)'s correction factor and further simplify it to include only the elements containing information with respect to the parameters of interest by

$$p(\beta|Y,D) \propto p(\beta)\cdot \prod_{n=1}^{N} \frac{\exp(V_{in}+ln(\pi(D_n|i)))}{\sum_{j\in D_n}\exp(V_{jn}+ln(\pi(D_n|j)))} \tag{7}$$

$$\propto p(\beta)\cdot \prod_{n=1}^{N} \frac{\exp(V_{in})}{\sum_{j\in D_n}\exp(V_{jn}+ln(\pi(D_n|j)))}, \tag{8}$$

it becomes directly clear that $\pi(D_n|i)$ only scales the likelihood function but is irrelevant for estimation purposes. Only $\pi(D_n|j)$ is relevant to correct for the loss of information with respect to the parameters of interest induced by the chosen sampling protocol.

3.1. Information divergence in the MNL posterior

The Kullback–Leibler divergence criterion (Kullback and Leibler, 1951) is frequently used in Bayesian statistics as a measure of information loss when approximating the 'true' distribution with an alternative distribution. Under the assumption of identical prior densities, the Kullback–Leibler divergence criterion (D_{KL}) reduces to the difference in the expected log-likelihood between the two models plus the log of the ratio of the marginal likelihoods as illustrated by[2]:

$$D_{KL} = \int_\beta p(\beta|Y,C) ln\left(\frac{p(\beta|Y,C)}{p(\beta|Y,D)}\right) d\beta = \int_\beta p(\beta|Y,C) ln\left(\frac{\frac{p(\beta)p(Y|C;\beta)}{p(Y|C)}}{\frac{p(\beta)p(Y|D;\beta)}{p(Y|D)}}\right) d\beta \tag{9}$$

$$D_{KL} = \int_\beta p(\beta|Y,C) ln\left(\frac{p(Y|C;\beta)}{p(Y|D;\beta)}\right) d\beta + ln\left(\frac{p(Y|D)}{p(Y|C)}\right). \tag{10}$$

The latter term is also known as the (inverse) *Bayes Factor*, which is independent of β and can accordingly be placed outside of the integral. The loss in information across the posterior can thus be separated into two elements, (i) the loss in information with respect to the parameters of interest β; and (ii) the difference in fit between the two models. Ideally, both types of information loss are minimised when using the sampled model, but for the purposes of estimation minimising only the first term is essential. Note that this term, $ln\left(\frac{p(Y|C;\beta)}{p(Y|D;\beta)}\right)$, is the negative of the information divergence studied by Keane and Wasi (2016) and covered in Appendix B and Section 2.3. It can be rewritten such that two familiar terms eventually emerge again:

$$ln\left(\frac{p(Y|C;\beta)}{p(Y|D;\beta)}\right) = \sum_{n=1}^{N} ln\left(\frac{exp(V_{in})}{\sum_{j\in C_n} exp(V_{jn})}\right) - ln\left(\frac{exp(V_{in}+ln(\pi(D_n|i)))}{\sum_{j\in D_n} exp(V_{jn}+ln(\pi(D_n|j)))}\right) \tag{11}$$

$$= \sum_{n=1}^{N} ln\left(\frac{\sum_{j\in D_n} exp(V_{jn}+ln(\pi(D_n|j)))}{\sum_{j\in C_n} exp(V_{jn})}\right) - \sum_{n=1}^{N} ln\left(\pi(D_n|i)\right) \tag{12}$$

$$= \sum_{n=1}^{N} ln\left(\sum_{j\in D_n} p(j|C_n;\beta)\cdot\pi(D_n|j)\right) - \sum_{n=1}^{N} ln\left(\pi(D_n|i)\right) \tag{13}$$

$$= \sum_{n=1}^{N} ln\left(\sum_{j\in C_n} p(j|C_n;\beta)\cdot\pi(D_n|j)\right) - \sum_{n=1}^{N} ln\left(\pi(D_n|i)\right) \tag{14}$$

$$= \sum_{n=1}^{N} ln\left(\pi(D_n|\beta)\right) - \sum_{n=1}^{N} ln\left(\pi(D_n|i)\right). \tag{15}$$

[2] The D_{KL} measure takes expectations with respect to β, but is conditional on observed choices and sampled choice sets.

For completeness, $\pi(D_n|\beta) = \sum_{j \in D_n} \pi(D_n|j)p(j|C_n;\beta)$ is the unconditional sampling probability of the smaller choice set $D_n \in C_n$. As argued in Section 2, $\sum_{n=1}^{N} ln(\pi(D_n|i))$ can be disregarded here because the term is independent of β. The only term which describes the (negative) information loss with respect to β is $\sum_{n=1}^{N} ln\left(\pi(D_n|\beta)\right)$ Following the same logic as applied in Appendix B, we now aim to minimise the expected information loss with respect to β across the entire posterior distribution (not just the point estimate). This corresponds to maximising $\mathbb{E}\left(\sum_{n=1}^{N} ln\left(\pi(D_n|\beta)\right)\right)$ since the sign of $\sum_{n=1}^{N} ln\left(\pi(D_n|\beta)\right)$ is always negative. Expectations are taken over all possible choices made, choice sets sampled, and all possible values of β. To this end, we define the joint probability of observing the vector of parameters β, the choice for alternative i and sampled choice set D_n by[3]:

$$\pi(\beta, i, D_n) = p(\beta) \cdot \pi(D_n|\beta) \cdot P(i|D_n;\beta). \tag{16}$$

$\mathbb{E}\left(\sum_{n=1}^{N} ln\left(\pi(D_n|\beta)\right)\right)$ is then given by:

$$\mathbb{E}\left(\sum_{n=1}^{N} ln\left(\pi(D_n|\beta)\right)\right) = \int_{\beta} p(\beta) \sum_{n=1}^{N} \sum_{D_n \in C_n} \sum_{i \in D_n} \pi(D_n|\beta) \cdot P(i|D_n;\beta) \cdot ln\left(\pi(D_n|\beta)\right) d\beta \tag{17}$$

$$= \int_{\beta} p(\beta) \sum_{n=1}^{N} \sum_{D_n \in C_n} \pi(D_n|\beta) ln(\pi(D_n|\beta)) d\beta. \tag{18}$$

Since entropy $\sum_{D_n \in C_n} \pi(D_n|\beta) ln(\pi(D_n|\beta))$ is maximised when applying McFadden (1978)'s correction term $ln(\pi(D_n|i))$, it minimises the expected information loss in the posterior at *every* value of β for a given sampling protocol. Therefore, the emerging posterior distribution using McFadden (1978)'s correction factor – either based on uniform or positive conditioning – provides the best approximation of the shape and location of the true posterior density in Bayesian MNL models.

3.2. Bayesian posterior densities and point estimates

The above results highlight that McFadden (1978)'s correction factor has good *ex-ante* finite and large sample properties because it minimises the expected information loss with respect to the parameters of interest under the sampling of alternatives irrespective of sample size. In other words, it provides the best approximation of the shape and location of the 'true' posterior distribution. As the sample size increases it can be shown that the Bayesian point estimate converges to the (consistent) point estimate obtained using classical maximum likelihood estimation.

Train (2009, chapter 12) highlights using the Bernstein–von Mises Theorem that the posterior mean is generally considered to be the best point estimate from the posterior density under many different loss functions. He furthermore illustrates that as the sample size increases (i) the posterior density converges to a Gaussian density, and (ii) the posterior variance becomes the same as that of the maximum likelihood estimate because the information in the data outweighs the information contained in the prior. Additionally, the posterior mean asymptotically converges to the maximum likelihood estimate such that asymptotically the classical sampling distribution and the Bayesian posterior distribution are the same. As a result, for all sampling protocols satisfying either a uniform or positive conditioning, McFadden (1978)'s correction factor will provide consistent parameter estimates in both the classical and Bayesian settings.

As the sample size decreases two effects occur. First, posterior and classical standard errors on the parameter estimates increase because there is less information in the data regarding β by default. Second, sampling of alternatives is likely to generate some additional degree of error because we sample fewer chosen alternatives (from the population) and choice sets D_n (when applying sampling of alternatives) across the sample. Only when the sample size becomes sufficiently large, the actual information loss with respect to β (in the parameter estimate and in the posterior) will converge in probability to the expected information loss (as per McFadden (1978)'s proof in Appendix A).

Indeed, one can correct for the actual loss in information in estimation due to sampling of alternatives, but this would require evaluating the full denominator of the MNL choice probability and defies the purpose of applying sampling of alternatives. McFadden (1978)'s correction protocol thus provides the best choice of correction factor without knowing which alternatives will be sampled a priori and irrespective of sample size. Determining whether sample sizes are sufficiently large, however, remains an empirical question and is no different for Bayesian and classical estimation. Similarly, determining the severity of the loss in information for the parameters of interest should be addressed empirically, including the search for optimal sampling rates (Daly et al., 2014).

4. Sampling of alternatives in Bayesian MMNL

We now switch our attention to the mixed multinomial logit (MMNL) model. MMNL assumes that the individual-level parameters β_n are distributed over the population of interest. The mixing density $f(\beta_n|\theta)$ describes the distribution of preferences, where θ is the set of hyper-parameters characterising the mixing density. For example, when β_n is assumed to follow a normal density then θ comprises the parameters characterising the mean and covariance matrix.

[3] We still assume that the sampling protocol is independent of β and without loss of generality that other explanatory variables X_n, and C_n are assumed to be non-stochastic. Moreover, we assume that choices and sampling probabilities are independent across choice observations.

4.1. Bayesian estimation of MMNL models and the role of data augmentation

Let Eq. (19) describe the likelihood function of the panel MMNL model, where β_n is assumed to be constant across observations by the same individual $t = 1, \ldots, T$. The model reduces to the cross-sectional MMNL model for $T = 1$. Estimating the parameters of interest θ can be done using Bayesian and classical estimation methods. Both estimation methods are extensively covered by Train (2009, see for example chapters 6, 10 and 12). Classical estimation typically approximates the (log of the) likelihood function

$$P(Y|\theta, C) = \prod_{n=1}^{N} \int \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n) f(\beta_n|\theta) d\beta_n \tag{19}$$

by taking a large number of draws from $f(\beta_n|\theta)$ and is accordingly referred to as a maximum simulated likelihood (MSL) approach. Bayesian estimation does not optimise the (same) likelihood function but requires simulation methods to characterise the posterior density for θ. Section 12.6 in Train (2009) provides an accessible introduction to the process of sequentially drawing from conditional posterior densities for the MMNL model — also known as Gibbs Sampling. Once the Gibbs Sampler (GS) has converged, the draws for the hyper-parameters θ describe its posterior density.

Assuming for illustration purposes only that the MMNL model only contains normally distributed random parameters with a mean μ and covariance matrix Σ, the GS comprises three steps repeated a large number of times and starting from some arbitrary values for $\beta_n \forall n$ and Σ (see Train, 2009, pp 301–302).[4]

1. $\mu|\beta_n \forall n, \Sigma$, conditional on values for Σ and $\beta_n \forall n$, update μ by taking a draw from $p(\mu|\beta_n, \Sigma) = f(\beta_n|\mu, \Sigma)p(\mu)$. Here $p(\mu|\beta_n, \Sigma)$ describes the conditional posterior for μ, the mixing density $f(\beta_n|\mu, \Sigma)$ is the 'likelihood' and $p(\mu)$ is the prior density on μ. Train (2009) shows that a multivariate normal mixing density $f(\beta_n|\mu, \Sigma)$ together with a multivariate normal prior density results in a normal posterior from which it is easy to take a draw.
2. $\Sigma|\beta_n \forall n, \mu$, using the new value for μ and conditional on $\beta_n \forall n$, update Σ by taking a draw from $p(\Sigma|\beta_n, \mu) = f(\beta_n|\mu, \Sigma)p(\Sigma)$. Here $p(\Sigma|\beta_n, \mu)$ describes the conditional posterior for Σ, the mixing density $f(\beta_n|\mu, \Sigma)$ is the 'likelihood' and $p(\Sigma)$ is the prior density on the covariance matrix. Train (2009) shows that a multivariate normal mixing density $f(\beta_n|\mu, \Sigma)$ together with an inverted Wishart normal prior density results in an inverted Wishart posterior from which it is easy to take draws. Akinc and Vandebroek (2018) discuss alternative priors that can be used without changing the three-step nature of the GS.
3. $\beta_n \forall n|\mu, \Sigma$, using the new values for μ and Σ, update each individual level parameter $\beta_n \forall n$ by taking a draw from $p(\beta_n|Y_n, C_n, \mu, \Sigma) = \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n) f(\beta_n|\mu, \Sigma)$. Here $p(\beta_n|Y_n, C_n, \mu, \Sigma)$ describes the conditional posterior for the individual level parameter for individual n, $P(Y_{nt}|C_{nt}; \beta_n)$ is the MNL probability of observing the choice made by individual n in choice task t, and $f(\beta_n|\mu, \Sigma)$ is the prior density of β_n. Given the presence of the MNL probability, it is generally impossible to find a prior (i.e. mixing density) that will result in a convenient shape for the posterior distribution for which it is easy to draw from. Train (2009, pp 302) describes how the Metropolis–Hastings algorithm can be used in this case to take suitable draws from the conditional posterior density.

Changes in the shape of the mixing density $f(\beta_n|\theta)$ do not alter the structure of the GS. It only affects the way in which the hyper-parameters are updated in Steps 1 and 2 and some of the calculations in the Metropolis–Hastings algorithm (see for example Blasi et al., 2010). The normal density used here is therefore not a special case, and the discussion below holds without loss of generality.

The three-step procedure in the GS explains the terminology of 'Hierarchical Bayes' often found in the Bayesian MMNL literature. That is, the mixing density acts as a prior on the individual-level parameter β_n in Step 3, whereas in Steps 1 and 2 the mixing density acts as the likelihood and an additional layer of prior densities is required on the hyper-parameters of the mixing density. This hierarchy emerges because the individual-level parameters are not integrated out – as happens in MSL – but they are actually estimated. Step 3 takes draws for β_n for each individual and at the end of the GS, the posterior for each individual-level parameter can be characterised with the stored draws. This process of estimating the latent parameters β_n is also known as data augmentation (Tanner and Wong, 1987). The computational benefit of augmenting β_n is that conditional on β_n, the choice probability is MNL which is easy to evaluate whilst avoiding the need for integration.

By implementing data augmentation, Bayesian MMNL models directly estimate the individual-level parameters alongside the hyperparameters of the mixing density resulting in the joint posteriors $p(\theta, \boldsymbol{\beta}^+|Y, C)$ and $p(\theta, \boldsymbol{\beta}^+|Y, D)$, where $\boldsymbol{\beta}^+$ comprises all individual-level parameters β_n across all individuals. The '+' is added to avoid confusion with β in the MNL model. Chan et al. (2019, Chapter 14 pp. 239) highlight that the draws from this joint posterior can be used to characterise the marginal posteriors $p(\theta|Y, C)$ and $p(\theta|Y, D)$. We will make extensive use of the latter property in deriving our results for the sampling of alternatives in Bayesian MMNL models. Our primary interest is in estimating θ, and not β_n, which saves a lot of computer memory by not storing the draws for β_n.

[4] The GS can easily be extended to include fixed parameters.

4.2. Bayesian estimation of MMNL models under sampling of alternatives

Returning back to the challenge of estimating choice models with large choice sets, note that the MNL choice probability is only part of Step 3 of the GS. Once a draw for $\beta_n \forall n$ is taken Steps 1 and 2 are not influenced by the choice set size. Hence, sampling of alternatives may be able to address the computational challenge arising in Step 3 of the GS by approximating the conditional posterior density $p(\beta_n|Y_n, C_n, \theta)$ by $p(\beta_n|Y_n, D_n, \theta)$.

All that needs to be recognised now is that the conditional posterior density $p(\beta_n|Y_n, C_n, \theta)$ has the same structure as the posterior derived for the MNL model but at the individual level. Since our results for MNL in Section 3 apply to any sample size, McFadden (1978)'s correction factor minimises the expected loss in information in the conditional posterior for the individual-level parameter β_n assuming that (i) the sampled choice set D_{nt} includes the chosen alternative, (ii) the sampling protocol satisfies positive conditioning. No additional information loss occurs in relation to the conditional posteriors for θ, i.e. steps 1 and 2 of the GS. Since the draws from the GS converge to the draws of the joint posterior (and can be used to characterise the marginal posterior density for θ), McFadden (1978)'s correction factor also minimises the expected information loss of the overall MMNL model.

Appendix C provides a formal proof that the expected loss of information with respect to the parameters of interest (β^* and θ) in the MMNL model under data augmentation are minimised when using McFadden (1978)'s correction factor for all sampling protocols satisfying either uniform or positive conditioning. This result is particularly encouraging as it enables researchers to combine the computational benefits of the sampling of alternatives with those of emerging computationally efficient Bayesian estimators. Namely, data augmentation – which is crucial in extending our results from MNL to MMNL – is universally applicable to other computationally efficient Bayesian estimators, such as variational Bayes (Bansal et al., 2020; Rodrigues, 2022).

4.3. Contrasting sampling of alternatives for MMNL with Bayesian and MSL methods

Bayesian posterior analysis does not rely on the asymptotic sampling distribution, the results for MNL and MMNL apply to samples of any size N. McFadden (1978)'s correction factor will minimise the expected information loss in the full posterior density with respect to the parameters of interest, whether that is β in MNL or θ in MMNL. This result applies to all sampling protocols satisfying either uniform or positive conditioning.

In relation to large sample sizes, we can again invoke the Bernstein-von Mises Theorem such that as the number of respondents N becomes sufficiently large the marginal posterior densities $p(\theta|Y, C)$ and $p(\theta|Y, D)$ converge to the asymptotic sampling distributions of their maximum likelihood counterparts.[5] MSL additionally requires the number of draws R to approximate the integral in the MMNL likelihood function to rise faster than $\sqrt{N}$ in order for classical estimation of the MMNL model to be "*consistent, asymptotically normal, efficient and equivalent to maximum likelihood*" (Train, 2009, Chapter 10, pp. 256). Since Bayesian estimation of the MMNL model using data augmentation does not approximate the referred integral the requirement on R does not transfer to the Bayesian setting.

We now focus on the theoretical need for an additional correction factor in classical MSL methods, as covered by Guevara and Ben-Akiva (2013a) and Keane and Wasi (2016). Following Guevara and Ben-Akiva (2013a), if we would know β_n - *which is the case under data augmentation in Bayesian estimation* - then the joint probability of observing the choice for alternative i and sampled choice set D_{nt} would be described by:

$$\pi(D_{nt}, i|\beta_n) = \pi(D_{nt}|\beta_n) \cdot P(i|D_{nt}; \beta_n). \tag{20}$$

MSL, however, requires integrating out the uncertainty regarding β_n. This results in the following expression conditional on the true hyper-parameters θ^* of the mixing density:

$$\pi(D_{nt}, i|\theta^*) = \int_{\beta_n} \pi(D_{nt}|\beta_n) \cdot P(i|D_{nt}; \beta_n) f(\beta_n|\theta^*) d\beta_n. \tag{21}$$

Following Bayes' rule we can accordingly define:

$$\pi(i|D_{nt}, \theta^*) = \frac{\pi(D_{nt}, i|\theta^*)}{\pi(D_{nt}|\theta^*)} \tag{22}$$

$$= \frac{\int_{\beta_n} \pi(D_{nt}|\beta_n) P(i|D_{nt}; \beta_n) f(\beta_n|\theta^*) d\beta_n}{\pi(D_{nt}|\theta^*)} \tag{23}$$

$$= \int_{\beta_n} W_{nt} P(i|D_{nt}; \beta_n) f(\beta_n|\theta^*) d\beta_n, \tag{24}$$

where:

$$W_{nt} = \frac{\pi(D_{nt}|\beta_n)}{\pi(D_{nt}|\theta^*)} = \frac{\sum_{j \in D_{nt}} \pi(D_{nt}|j) P(j|C_{nt}; \beta_n)}{\sum_{j \in D_{nt}} \pi(D_{nt}|j) P(j|C_{nt}; \theta^*)}. \tag{25}$$

Guevara and Ben-Akiva (2013a) show that consistent estimates are obtained when in addition to McFadden's correction factor the term W_{nt} is included in the corrected likelihood function. In short, the latent nature of β_n in MSL (and latent class modelling) introduces the theoretical need for including an additional correction factor. Since data augmentation in Bayesian estimation resolves

[5] Blasi et al. (2010) proof the consistency of Bayesian MMNL models, based on the described GS structure, and their results apply to a wide variety of mixing densities, including non-parametric ones.

the latent nature of β_n, there is no longer the theoretical need for the additional correction factor W_{nt}. Note that Bayesian estimation without data augmentation requires the inclusion of W_{nt}. We return to this in Section 5.

Moving to the empirical application of W_{nt}, Guevara and Ben-Akiva (2013a) highlight the dependency of W_{nt} on the full choice set. To circumvent this issue, they develop a feasible and consistent estimator approximating W_{nt} only using elements of the sampled choice set D_{nt}. Three approaches are proposed, respectively using (i) population shares, (ii) observed choices by the individual and (iii) the naive method which sets W_{nt} to 1 and thereby reduces to applying only McFadden (1978)'s correction factor. Since the naive approximation of W_{nt} provides consistent estimates and provides good results in Monte Carlo simulations and real-world examples, this is the recommended approach by Guevara and Ben-Akiva (2013a).

Von Haefen and Domanski (2018) reach a similar conclusion in the context of applying sampling of alternatives in latent class modelling in combination with uniform conditioning. In their case, an alternative explanation is, however, appropriate. Namely, they implement the Expectations–Maximisation (EM)-algorithm. Instead of integrating out the uncertainty regarding class membership, which is the traditional approach to estimating latent class models, the EM-algorithm sequentially estimates weighted class specific MNL models (Bhat, 1997; Train, 2009). Since in these weighted class specific MNL models there are no longer latent variables present, McFadden (1978)'s correction factor is therefore sufficient for obtaining consistent parameter estimates. Where the EM-algorithm determines the weights based on posterior probabilities for the class membership, the Bayesian alternative would be the augmentation of class membership, again removing the theoretical need for the implementation of an additional correction factor.

In effect, the Bayesian, the EM and the MSL approach argue that McFadden (1978) correction factor is the only necessary correction factor for applying sampling of alternatives in MMNL and latent class models. To arrive at this conclusion, both approaches take a different avenue. Bayesian estimation circumvents the problem of latent β_n or class membership by augmenting the parameter and directly estimating it negating the need for an additional correction factor. The EM algorithm breaks down the estimation into sequentially estimating simpler weighted MNL models where McFadden (1978)'s correction factor is sufficient. Classical estimation methods, however, acknowledge the latent nature of β_n and argue that in principle an additional correction factor is required. In practice, the need for this additional correction factor is negligible. Keane and Wasi (2016), for example, show in a Monte Carlo analysis that the bias introduced is very limited for modest-sized subsets D_{nt}. If the number of respondents becomes sufficiently large all three estimation approaches will result in consistent parameter estimates for *any* sampling protocol satisfying positive conditioning when McFadden (1978)'s correction factor is applied.

5. Monte Carlo analyses

This section presents two sets of Monte Carlo analyses illustrating the implementation of sampling of alternatives using Bayesian and Classical estimators of MMNL. The first Monte Carlo analysis adopts the naive sampling approach for both types of estimators and contrasts their outcomes across different samples sizes, number of choices made per respondent, choice set sizes, and number of alternatives sampled. The second Monte Carlo analysis explores the inclusion of Guevara and Ben-Akiva (2013a)'s additional correction factor in Bayesian estimation and contrasts the results with the naive approach using data augmentation advocated here.

5.1. Performance of the naive approach across Bayesian and classical estimators

Table 1 summarises the implemented simulation settings for the first Monte Carlo analysis. We assume that a group of 250 or 1000 individuals (N) is making a sequence of either 5 or 10 choices (T) each. The number of alternatives in each choice set C_{nt} is 50 or 100, from which respectively 20 or 10, or 30 or 10 alternatives are sampled (including the chosen alternative). The data generating process is repeated such that for each of the 16 unique combinations of settings, 30 datasets are generated and the corresponding models are estimated. The resampling happens at the level of the choice data, not the sampling of alternatives. Both classical and Bayesian estimation of MMNL is done using the true and sampled choice sets. In the Bayesian estimation, non-informative priors are used and we take 20,000 posterior draws (burn-in: 10,000, thinning:10, effective posterior draws: L = 1000), which are sufficient for convergence as the Gelman–Rubin Diagnostic is close to 1 for all parameters. For the MSL approach, 100 draws were taken using modified Latin hypercube sampling to approximate the integral in the likelihood function (Hess et al., 2006).

Each alternative in the choice set is characterised by four attributes. Following Keane and Wasi (2016), the first two attributes are drawn from standard normal distributions and are associated with normally distributed random parameters with mean {1;1} and covariance matrix $\begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$. The third and fourth attributes are dummy variables where the value of one is associated with a probability of 50%. These two dummy variables are associated with fixed parameter values of {1;-1}. The sampled choice set includes the chosen alternative and the remainder of the sampled choice set is obtained using uniform sampling without replacement such that we are working in the context of uniform conditioning and no correction factor is required in practice.

We summarise the results of the simulation study by computing the following five metrics: (a) mean of the parameter estimates across the 30 repetitions; (b) standard deviation around this mean ; (c) average absolute percentage bias (APB); (d) average coverage probability (CP) in percentage; and (e) mean of the standard error of the parameter estimates across the 30 repetitions. Using the classical estimation setting, as an example, we compute the APB and the CP of a parameter corresponding to a synthetic dataset as follows:

$$APB = \frac{100}{R}\sum_{r=1}^{R}\left|\frac{\hat{\beta}_r - \beta^*}{\beta^*}\right| \tag{26}$$

Table 1
Monte Carlo analysis setup.

	Smaller choice set size	Larger choice set size
Choice set size (C_{nt})	50	100
Size of sampled subset (D_{nt})	{20,10}	{30,10}
Number of individuals (N)	{250,1000}	{250,1000}
Choice tasks per individual (T)	{5,10}	{5,10}
Number of MC resamples	30	30
Estimation	Classical and Bayesian MMNL	Classical and Bayesian MMNL

Table 2
True choice set size 50, no sampling of alternatives applied.

N	T	True		Classical — full choice set					Bayesian — full choice set				
				Mean	SD	APB	CP	Mean st.er.	Mean	SD	APB	CP	Mean st.er.
250	5	B1	1.00	0.98	0.08	6.38%	93.33%	0.07	1.00	0.07	5.84%	93.33%	0.07
		B2	1.00	0.97	0.09	7.50%	93.33%	0.07	0.99	0.08	6.33%	86.67%	0.07
		B3	1.00	1.02	0.03	2.85%	93.33%	0.04	1.02	0.03	3.05%	93.33%	0.04
		B4	−1.00	−1.01	0.03	2.56%	90.00%	0.03	−1.01	0.03	2.59%	93.33%	0.04
		C1	0.60	0.60	0.08	13.86%	93.33%	0.06	0.57	0.10	13.72%	93.33%	0.10
		C2	1.00	0.98	0.06	9.67%	96.67%	0.08	0.98	0.11	10.76%	96.67%	0.13
		C3	1.00	0.96	0.07	11.08%	96.67%	0.06	0.98	0.13	8.62%	93.33%	0.13
	10	B1	1.00	0.97	0.08	6.63%	86.67%	0.07	0.99	0.06	4.89%	96.67%	0.07
		B2	1.00	0.97	0.07	6.07%	86.67%	0.07	0.99	0.06	4.47%	93.33%	0.07
		B3	1.00	1.00	0.02	1.66%	90.00%	0.02	1.01	0.02	1.69%	93.33%	0.03
		B4	−1.00	−1.00	0.02	1.86%	93.33%	0.02	−1.00	0.02	1.80%	100.00%	0.03
		C1	0.60	0.64	0.06	12.59%	93.33%	0.05	0.60	0.08	16.93%	96.67%	0.09
		C2	1.00	1.04	0.05	8.63%	86.67%	0.07	1.00	0.09	13.31%	100.00%	0.11
		C3	1.00	1.03	0.06	12.08%	93.33%	0.05	1.01	0.11	7.19%	96.67%	0.11
1000	5	B1	1.00	0.98	0.04	5.40%	96.67%	0.04	1.01	0.07	5.87%	96.67%	0.04
		B2	1.00	0.98	0.04	5.89%	86.67%	0.04	1.00	0.07	5.82%	93.33%	0.04
		B3	1.00	0.99	0.02	2.41%	90.00%	0.02	1.00	0.03	2.41%	96.67%	0.02
		B4	−1.00	−1.00	0.02	1.96%	86.67%	0.02	−1.00	0.02	1.90%	100.00%	0.02
		C1	0.60	0.65	0.05	12.84%	100.00%	0.03	0.62	0.08	9.63%	96.67%	0.05
		C2	1.00	1.02	0.09	10.84%	93.33%	0.04	1.02	0.14	11.51%	93.33%	0.07
		C3	1.00	1.05	0.04	12.08%	96.67%	0.03	1.05	0.13	10.51%	96.67%	0.07
	10	B1	1.00	0.98	0.04	7.30%	86.67%	0.03	0.98	0.08	6.07%	93.33%	0.03
		B2	1.00	0.98	0.04	7.57%	90.00%	0.03	0.99	0.08	5.98%	93.33%	0.03
		B3	1.00	1.00	0.01	1.79%	93.33%	0.01	1.00	0.02	1.83%	96.67%	0.01
		B4	−1.00	−1.00	0.01	1.85%	96.67%	0.01	−1.00	0.02	1.87%	93.33%	0.01
		C1	0.60	0.62	0.05	15.89%	100.00%	0.03	0.61	0.08	10.61%	100.00%	0.05
		C2	1.00	1.03	0.03	11.74%	90.00%	0.03	1.02	0.10	9.71%	96.67%	0.06
		C3	1.00	1.05	0.03	11.87%	100.00%	0.02	1.05	0.11	8.13%	90.00%	0.06

N: respondents; T: choice tasks; Mean: avg. estimate across 30 MC resamples; SD: St dev of estimate across 30 MC resamples.
APB: average percentage bias; CP: coverage probability; Mean st.er.: average of standard error across 30 MC resamples.

$$CP = \frac{100}{R}\sum_{r=1}^{R} I\left[\hat{\beta} - 1.96 \cdot st.error(\hat{\beta}) \leq \beta^{*} \leq \hat{\beta} + 1.96 \cdot st.error(\hat{\beta})\right] \tag{27}$$

where $I[\cdot]$ is the indicator function and R is the number of synthetic datasets which is 30 in our case. Note that the corresponding values in the Bayesian setting relate to the posterior mean and posterior standard deviation in the APB computation. To compute CP in the Bayesian estimation, we compute the percentage of synthetic datasets where the true value lies in the 95% highest posterior density interval.

Table 2 presents the results assuming a choice set of 50 alternatives whilst estimating standard MMNL models without a sampling of alternatives using classical and Bayesian methods. Overall, the true parameter estimates are recovered with a high level of accuracy across the 30 MC resamples. When the sample size increases either by increasing the number of respondents or the number of choices per respondent we generally see a reduction in the variability of the parameter estimates, their bias, increases in the average coverage probability and reductions in the average standard error of the parameter estimates. Table 2 confirms that for the base scenario, classical and Bayesian estimation methods have a comparable performance, and can therefore act as a point of reference in the context of sampling of alternatives applied below.

Tables 3 and 4 present the same set of results when respectively 20 and 10 alternatives are sampled from the full set of 50 alternatives. In both cases, we observe that the true parameters can be recovered but in general, the level of precision is lower for the sampled choice sets than the full choice set. This is to be expected because of the reduced level of information about the parameters of interest in each choice task as a result of evaluating a reduced set of alternatives. Consequently, (on average) the degree of variation in the parameter estimates increases, the average percentage bias increases, average coverage probabilities

Table 3
True choice set size 50, 20 alternatives sampled.

N	T	True		Classical — 20 sampled alts					Bayesian — 20 sampled alts				
				Mean	SD	APB	CP	Mean st.er.	Mean	SD	APB	CP	Mean st.er.
250	5	B1	1.00	1.01	0.08	6.53%	90.00%	0.07	1.02	0.08	6.78%	93.33%	0.08
		B2	1.00	1.01	0.09	6.70%	83.33%	0.08	1.02	0.09	6.39%	86.67%	0.08
		B3	1.00	1.04	0.04	4.80%	86.67%	0.04	1.04	0.04	5.07%	86.67%	0.04
		B4	−1.00	−1.03	0.04	3.86%	93.33%	0.04	−1.03	0.04	4.16%	93.33%	0.04
		C1	0.60	0.61	0.08	14.48%	93.33%	0.07	0.60	0.10	13.95%	96.67%	0.11
		C2	1.00	0.99	0.07	10.61%	100.00%	0.09	1.07	0.13	9.90%	96.67%	0.15
		C3	1.00	1.01	0.08	12.61%	93.33%	0.07	1.08	0.17	13.19%	90.00%	0.15
	10	B1	1.00	0.99	0.07	5.07%	90.00%	0.07	1.01	0.07	5.46%	96.67%	0.07
		B2	1.00	1.00	0.07	4.88%	96.67%	0.07	1.01	0.06	4.83%	90.00%	0.07
		B3	1.00	1.01	0.03	2.40%	93.33%	0.03	1.02	0.03	2.54%	93.33%	0.03
		B4	−1.00	−1.01	0.03	2.03%	93.33%	0.03	−1.01	0.03	2.10%	96.67%	0.03
		C1	0.60	0.61	0.08	14.05%	90.00%	0.06	0.62	0.10	13.21%	96.67%	0.10
		C2	1.00	1.02	0.29	7.56%	90.00%	0.07	1.03	0.10	7.71%	93.33%	0.13
		C3	1.00	1.04	0.07	12.21%	93.33%	0.05	1.05	0.14	12.04%	96.67%	0.12
1000	5	B1	1.00	1.01	0.04	6.21%	93.33%	0.04	1.04	0.07	6.83%	100.00%	0.04
		B2	1.00	1.01	0.04	6.69%	93.33%	0.04	1.04	0.07	6.86%	96.67%	0.04
		B3	1.00	1.02	0.02	3.24%	83.33%	0.02	1.02	0.04	3.49%	96.67%	0.02
		B4	−1.00	−1.02	0.02	3.09%	86.67%	0.02	−1.02	0.03	3.22%	93.33%	0.02
		C1	0.60	0.68	0.05	17.30%	86.67%	0.04	0.67	0.11	14.47%	93.33%	0.06
		C2	1.00	1.06	0.09	14.19%	90.00%	0.04	1.09	0.18	13.45%	86.67%	0.08
		C3	1.00	1.09	0.04	15.46%	93.33%	0.03	1.11	0.16	15.11%	86.67%	0.08
	10	B1	1.00	1.00	0.04	6.03%	90.00%	0.03	1.00	0.07	5.64%	90.00%	0.04
		B2	1.00	1.00	0.03	6.51%	96.67%	0.03	1.01	0.08	6.08%	93.33%	0.04
		B3	1.00	1.01	0.01	2.22%	83.33%	0.01	1.01	0.03	2.23%	93.33%	0.01
		B4	−1.00	−1.01	0.01	2.53%	90.00%	0.01	−1.02	0.03	2.56%	90.00%	0.01
		C1	0.60	0.65	0.05	14.00%	83.33%	0.03	0.62	0.08	11.20%	100.00%	0.05
		C2	1.00	1.06	0.03	9.41%	80.00%	0.04	1.04	0.11	8.72%	93.33%	0.06
		C3	1.00	1.07	0.03	10.84%	76.67%	0.03	1.07	0.11	10.49%	93.33%	0.06

N: respondents; T: choice tasks; Mean: avg. estimate across 30 MC resamples; SD: St dev of estimate across 30 MC resamples.
APB: average percentage bias; CP: coverage probability; Mean st.er.: average of standard error across 30 MC resamples.

decrease and the mean standard errors increase across the 30 MC resamples. This effect becomes more pronounced when sampling 10 instead of 20 alternatives. Especially in the latter case, we observe some CP values decrease to under 70%, where good CP values are considered to be above 85% and especially the average percentage bias is approaching 20% for some of the parameters in the covariance matrix. These results are consistent between classical and Bayesian estimation. In this case, we would recommend that sampling 10 alternatives (20%) is too few. One of the ways in which Bayesian estimation may circumvent such issues is when additional information is available and included through the use of informed prior densities, but we have not employed this strategy here.

Table 5 presents the results for the Bayesian analysis related to the setting with 100 alternatives. The results for classical estimation are similar and available in Appendix D. Again, the model using the full choice set is able to recover true parameters with relatively low levels of bias and acceptable levels of the average coverage probability. When sampling 30 out of 100 alternatives, we can see the levels for APB increase and the CP fall to lower levels across the different settings. At this sampling rate, the results are close to being acceptable, but when reducing the sampling to 10 out of 100 alternatives, we can clearly see CP levels falling across the board to unacceptable levels indicating that Nerella and Bhat (2004) suggestion of using a sampling rate of around 25% also holds in this context when accounting for covariances across the random parameters. It is interesting to see that across the models presented the trend is consistent irrespective of sample size. Indeed, there is some additional bias and estimation imprecision in the context of the smaller sample sizes, but no specific trend emerges that in case of a decrease in sample sizes, higher sampling rates need to be used for sampling alternatives to be successful. We take this as supporting evidence of our theoretical result that McFadden (1978)'s correction factor also has desirable finite sample properties.

5.2. Performance of the additional correction factor in a Bayesian approach

One of our key arguments is that due to data augmentation in Bayesian MMNL there is no longer the *theoretical* need to include the additional correction factor W_{nt} in estimation. In fact, if one were to implement W_{nt} in Bayesian estimation the structure of the posterior density changes, because W_{nt} includes an integral *and* depends on the parameters (mean and variance) of the (normal) mixing density. As such, implementing data augmentation is no longer a feasible strategy to simplify and increase the speed of estimation. The only feasible approach is to work with the quasi-likelihood described by Guevara and Ben-Akiva (2013a) and directly estimate the model including W_{nt} using a Metropolis–Hastings algorithm simulating the integral over all possible values of β_n.

Table 4
True choice set size 50, 10 alternatives sampled.

N	T	True		Classical — full choice set					Bayesian — full choice set				
				Mean	SD	APB	CP	Mean st.er.	Mean	SD	APB	CP	Mean st.er.
250	5	B1	1.00	1.04	0.10	8.49%	86.67%	0.08	1.05	0.09	8.84%	86.67%	0.09
		B2	1.00	1.03	0.10	7.75%	90.00%	0.08	1.04	0.10	8.41%	90.00%	0.09
		B3	1.00	1.07	0.05	7.19%	73.33%	0.05	1.07	0.05	7.79%	66.67%	0.05
		B4	−1.00	−1.06	0.05	6.77%	70.00%	0.05	−1.07	0.05	7.29%	70.00%	0.05
		C1	0.60	0.63	0.10	18.89%	90.00%	0.08	0.63	0.11	17.30%	93.33%	0.13
		C2	1.00	1.03	0.41	11.02%	96.67%	0.10	1.02	0.14	11.43%	96.67%	0.18
		C3	1.00	1.04	0.09	14.40%	86.67%	0.08	1.03	0.16	15.71%	100.00%	0.18
	10	B1	1.00	1.00	0.07	5.98%	90.00%	0.07	1.03	0.07	6.17%	93.33%	0.08
		B2	1.00	1.01	0.07	4.99%	93.33%	0.07	1.03	0.07	5.26%	90.00%	0.08
		B3	1.00	1.03	0.04	3.95%	83.33%	0.03	1.04	0.04	4.18%	83.33%	0.03
		B4	−1.00	−1.03	0.03	3.28%	90.00%	0.03	−1.03	0.03	3.47%	90.00%	0.03
		C1	0.60	0.68	0.07	16.32%	90.00%	0.06	0.64	0.09	13.68%	96.67%	0.11
		C2	1.00	1.10	0.06	12.52%	86.67%	0.08	1.08	0.10	10.33%	96.67%	0.14
		C3	1.00	1.09	0.07	12.44%	93.33%	0.06	1.07	0.14	11.94%	100.00%	0.14
1000	5	B1	1.00	1.03	0.04	7.88%	90.00%	0.04	1.06	0.08	8.52%	93.33%	0.04
		B2	1.00	1.03	0.05	8.98%	80.00%	0.04	1.06	0.09	9.64%	93.33%	0.04
		B3	1.00	1.04	0.03	6.23%	53.33%	0.02	1.06	0.04	6.67%	86.67%	0.02
		B4	−1.00	−1.04	0.02	5.17%	63.33%	0.02	−1.05	0.04	5.68%	86.67%	0.02
		C1	0.60	0.66	0.07	18.11%	80.00%	0.04	0.68	0.14	19.16%	93.33%	0.07
		C2	1.00	1.04	0.08	14.10%	80.00%	0.05	1.10	0.17	14.82%	86.67%	0.09
		C3	1.00	1.10	0.05	13.95%	96.67%	0.04	1.15	0.18	17.78%	93.33%	0.09
	10	B1	1.00	1.02	0.04	7.28%	93.33%	0.04	1.02	0.08	6.63%	90.00%	0.04
		B2	1.00	1.01	0.04	6.97%	93.33%	0.04	1.04	0.08	6.63%	93.33%	0.04
		B3	1.00	1.02	0.02	4.85%	63.33%	0.02	1.05	0.04	5.12%	66.67%	0.02
		B4	−1.00	−1.03	0.02	4.34%	63.33%	0.02	−1.05	0.03	4.71%	80.00%	0.02
		C1	0.60	0.67	0.05	21.24%	76.67%	0.03	0.66	0.12	18.61%	93.33%	0.05
		C2	1.00	1.10	0.03	15.95%	86.67%	0.04	1.11	0.16	14.81%	80.00%	0.07
		C3	1.00	1.10	0.03	14.64%	90.00%	0.03	1.12	0.16	15.21%	83.33%	0.07

N: respondents; T: choice tasks; Mean: avg. estimate across 30 MC resamples; SD: St dev of estimate across 30 MC resamples.
APB: average percentage bias; CP: coverage probability; Mean st.er.: average of standard error across 30 MC resamples.

In this subsection, we replicate the first Monte Carlo (random coefficients) experiment presented in Guevara and Ben-Akiva (2013a). In this case, $N = 1000$. $T = 1$, and $J = 1000$. Each alternative is characterised by a single attribute that is distributed Uniform (−2,1) for the first 500 alternatives and Uniform (−1, 2) for the second half. The corresponding parameter is assumed to be normally distributed with $\mu = 1.5$ and $\sigma = 0.8$. When implementing sampling of alternatives the sampled choice sets include respectively 5,30, and 50 alternatives using random sampling without replacement. The experiment is repeated 100 times ($R = 100$).

For the case of 5 sampled alternatives in the choice set we present three sets of results in Table 6. First, we report the original outcomes from Guevara and Ben-Akiva (2013a). Note that we only focus on the population shares and naive approaches due to the bad performance of the '1_0' method in the original paper. Second, we replicate the original experiment and present our corresponding classical estimations results. The small differences indicate we have been able to replicate the original simulation setting sufficiently close. Third, we present the same results using Bayesian estimation. The first two Bayesian models, respectively using the population shares and naive approaches both without data augmentation, are equivalent to the classical models and approximated W_{nt} during estimation. The final Bayesian model is our advocated modelling approach using the naive approach in combination with data augmentation. All Bayesian models display a reduction in bias for the model parameter estimates relative to their classical counterparts. This effect is, however, somewhat more pronounced when using our preferred approach using data augmentation. The MSE's are, however, comparable across the models and only small reductions are observed using our preferred approach. Across the specifications, and as expected, the t-test statistic indicates that in all cases the mean of the estimates across the 100 resamples cannot be differentiated from their true values. The count statistic confirms the improved performance of the naive approach over the population shares as the number is closer to 75 indicating that its empirical distribution is closer to its theoretical sampling distribution. Overall, there is however little difference in the performance of the different estimators, and together with the reduction in estimation time for the Bayesian model using data augmentation, this remains our recommended approach.

For completeness we present in Table 7 the results from increasing the size of the sampled choice set from 5 alternatives to 30 and 50 using classical estimation and our preferred Bayesian approach using data augmentation. Again, we observe comparable performance of the naive approaches and reductions in the size of the bias when the number of sampled alternatives increases.

6. Conclusions

In this paper, we have revisited McFadden (1978)'s correction factor for the sampling of alternatives. Our analysis has gone beyond the well-known result that the correction factor results in consistent parameter estimates in the context of multinomial

Table 5
True choice set size 100.

N	T	True		Bayesian — full choice set					Bayesian — 30 alts sampled					Bayesian — 10 alts sampled				
				Mean	SD	APB	CP	Mean st.er.	Mean	SD	APB	CP	Mean st.er.	Mean	SD	APB	CP	Mean st.er.
250	5	B1	1.00	1.00	0.04	3.40%	100.00%	0.08	1.02	0.04	3.73%	90.00%	0.09	1.04	0.04	4.59%	86.67%	0.08
		B2	1.00	0.99	0.04	3.31%	96.67%	0.07	1.02	0.04	3.40%	93.33%	0.09	1.04	0.05	4.83%	83.33%	0.07
		B3	1.00	1.00	0.02	1.67%	90.00%	0.03	1.02	0.02	2.25%	80.00%	0.05	1.04	0.03	4.47%	53.33%	0.04
		B4	−1.00	−1.00	0.02	1.56%	96.67%	0.03	−1.02	0.02	2.37%	83.33%	0.05	−1.04	0.02	4.49%	60.00%	0.04
		C1	0.60	0.60	0.07	8.73%	83.33%	0.10	0.64	0.07	10.13%	90.00%	0.14	0.66	0.09	13.76%	83.33%	0.12
		C2	1.00	1.00	0.07	6.65%	93.33%	0.13	1.04	0.08	7.47%	83.33%	0.19	1.07	0.09	9.25%	83.33%	0.16
		C3	1.00	1.01	0.08	5.42%	93.33%	0.13	1.06	0.09	8.82%	90.00%	0.18	1.09	0.11	10.83%	90.00%	0.15
	10	B1	1.00	1.00	0.03	2.61%	100.00%	0.07	1.02	0.03	3.14%	96.67%	0.08	1.04	0.04	4.29%	86.67%	0.08
		B2	1.00	1.00	0.03	2.73%	96.67%	0.07	1.02	0.03	2.98%	93.33%	0.08	1.03	0.04	3.96%	86.67%	0.07
		B3	1.00	1.00	0.01	1.22%	90.00%	0.02	1.01	0.01	1.26%	90.00%	0.03	1.02	0.02	2.48%	66.67%	0.04
		B4	−1.00	−1.00	0.01	0.93%	93.33%	0.02	−1.01	0.01	1.61%	86.67%	0.03	−1.03	0.02	2.83%	60.00%	0.04
		C1	0.60	0.60	0.04	5.62%	93.33%	0.09	0.62	0.05	6.92%	90.00%	0.11	0.64	0.05	7.85%	90.00%	0.10
		C2	1.00	1.01	0.06	3.45%	90.00%	0.11	1.04	0.07	6.24%	100.00%	0.15	1.06	0.08	7.74%	90.00%	0.12
		C3	1.00	1.00	0.05	4.98%	96.67%	0.11	1.04	0.05	5.38%	90.00%	0.15	1.06	0.05	6.76%	76.67%	0.12
1000	5	B1	1.00	1.00	0.03	2.44%	100.00%	0.04	1.02	0.03	3.03%	90.00%	0.04	1.05	0.04	5.05%	73.33%	0.04
		B2	1.00	1.00	0.03	2.68%	96.67%	0.04	1.02	0.04	3.09%	93.33%	0.04	1.05	0.04	4.82%	83.33%	0.03
		B3	1.00	1.00	0.01	1.04%	96.67%	0.02	1.02	0.02	2.28%	83.33%	0.03	1.07	0.02	6.53%	26.67%	0.02
		B4	−1.00	−1.00	0.02	1.45%	90.00%	0.02	−1.02	0.02	2.14%	80.00%	0.03	−1.05	0.03	5.41%	43.33%	0.01
		C1	0.60	0.59	0.05	7.56%	96.67%	0.05	0.62	0.05	8.21%	96.67%	0.07	0.65	0.06	11.07%	96.67%	0.06
		C2	1.00	1.01	0.06	3.93%	90.00%	0.06	1.05	0.07	6.90%	100.00%	0.09	1.09	0.09	11.29%	90.00%	0.07
		C3	1.00	1.00	0.05	4.81%	100.00%	0.06	1.04	0.05	5.67%	93.33%	0.09	1.08	0.07	8.59%	90.00%	0.07
	10	B1	1.00	1.00	0.04	3.25%	86.67%	0.03	1.02	0.04	3.53%	93.33%	0.04	1.03	0.04	4.43%	83.33%	0.04
		B2	1.00	1.00	0.03	2.33%	96.67%	0.03	1.02	0.03	2.67%	96.67%	0.04	1.04	0.03	4.08%	86.67%	0.03
		B3	1.00	1.00	0.01	0.91%	100.00%	0.01	1.02	0.01	1.67%	83.33%	0.02	1.04	0.01	3.80%	36.67%	0.02
		B4	−1.00	−1.00	0.01	0.98%	100.00%	0.01	−1.01	0.01	1.41%	86.67%	0.02	−1.04	0.02	3.84%	40.00%	0.01
		C1	0.60	0.60	0.05	7.37%	93.33%	0.04	0.62	0.06	8.80%	86.67%	0.06	0.64	0.07	10.63%	80.00%	0.05
		C2	1.00	1.01	0.07	4.98%	90.00%	0.06	1.04	0.08	6.91%	90.00%	0.07	1.06	0.08	8.23%	80.00%	0.06
		C3	1.00	1.01	0.06	5.61%	100.00%	0.06	1.04	0.06	5.70%	73.33%	0.07	1.07	0.07	7.71%	80.00%	0.06

N: respondents; T: choice tasks; Mean: avg. estimate across 30 MC resamples; SD: St dev of estimate across 30 MC resamples.
APB: average percentage bias; CP: Cerage probability; Mean st.er.: average of standard error across 30 MC resamples.

Table 6
Sampling 5 alternatives from $J = 1000$ using different estimators and R = 100.

Method	Stat.	Original results GBA (2013a)				Classical estimation				Bayesian estimation			
		Bias	MSE	*t-Test*	Count	Bias	MSE	*t-Test*	Count	Bias	MSE	*t-Test*	Count
Pop shares	μ	−0.09	0.016	1.07	56	−0.10	0.017	1.26	50	−0.08	0.012	0.90	60
	σ	−0.08	0.034	0.49	75	−0.11	0.037	0.65	70	−0.08	0.029	0.53	69
Naive	μ	−0.04	0.013	0.38	73	−0.03	0.010	0.35	72	−0.03	0.012	0.27	75
	σ	−0.12	0.039	0.73	71	−0.11	0.032	0.78	71	−0.11	0.036	0.69	70
Naive (data aug.)	μ									−0.02	0.011	0.21	69
	σ									−0.09	0.021	0.76	61

Bias: difference between the average estimator and the true value of each parameter.
Mean Squared Error (MSE): sum of the sampling variance and the square of the bias.
t-Test: ratio between the absolute value of the bias and the sampling standard deviation of the estimators.
Count: Calculated as the number of times the estimator of each repetition is within a 75% confidence interval of the true value.

Table 7
Sampling 30 and 50 alternatives from $J = 1000$ using different estimators and $R = 100$.

	Method	Stat.	30 alternatives sampled				50 alternatives sampled			
			Bias	MSE	t-Test	Count	Bias	MSE	t-Test	Count
Classical	Pop	μ	−0.061	0.007	1.10	41	−0.046	0.006	0.74	60
	shares	σ	−0.026	0.009	0.29	71	−0.022	0.011	0.22	76
	Naïve	μ	−0.007	0.005	0.09	72	−0.004	0.004	0.06	82
		σ	−0.020	0.010	0.21	73	−0.021	0.009	0.22	72
Bayesian	Naïve	μ	−0.007	0.005	0.10	70	−0.003	0.005	0.05	67
	(data aug)	σ	−0.032	0.010	0.34	72	−0.014	0.007	0.17	77

logit (MNL) models. The relation between the correction factor and the expected information loss with respect to the parameters of interest has been the centre of our attention. Building on the work of Keane and Wasi (2016), we have shown that for both uniform

and positive conditioning this expected loss of information – which is relevant for estimation purposes – is minimised at the true parameter values.

We provided an intuitive explanation for the source of information loss with respect to the parameters of interest. Since the sampling of alternatives evaluates a smaller number of utility differences in the denominator of the MNL choice probability, less information about the parameters of interest is obtained. By selecting a particular sampling protocol bias may arise in the parameter estimates by systematically influencing the subset of alternatives against which the chosen alternative is most likely contrasted in estimation. McFadden (1978)'s correction factor accounts for the latter effect. Because uniform conditioning does not steer the sampling process to a specific set of alternatives it makes intuitive sense that McFadden (1978)'s correction factor cancels out and is not required.

We have furthermore shown that this (expected) information loss with respect to the parameters of interest is an integral part of the Kullback–Leibler divergence criterion frequently used in Bayesian statistics measuring the loss of information between the true posterior density and an alternative approximation — based on the MNL likelihood under the sampling of alternatives in our case. In fact, we argue that for estimation purposes – which aims to learn about the parameters of interest – this is the only relevant term to consider. Accordingly, we were able to establish that McFadden (1978)'s correction factor minimises the expected loss in information with respect to the parameters of interest at every possible parameter value and therefore across the entire posterior density. The Bayesian MNL posterior based on McFadden (1978)'s correction factor under the sampling of alternatives is therefore the best approximation of the true posterior – in terms of minimum expected information loss – *irrespective* of sample size. As sample sizes decrease, the amount of information in the true and sampled model reduces and the degree of uncertainty increases (i.e. increased standard errors and bias in the parameter estimates). This happens irrespective of using a Bayesian or classical maximum likelihood approach. The only way by which Bayesian models could counteract such effects is by increasing the information content of the prior, i.e. by making use of informative priors. We have furthermore established that as the sample size increases the corresponding Bayesian point estimate will be consistent. McFadden (1978)'s correction factor therefore has desirable finite and large sample properties. The fact that sampling of alternatives transfers to Bayesian estimation methods together with desirable finite sample performance is an important contribution to the literature.

We continued our analysis by arguing that these convenient properties directly transfer to Bayesian MMNL models when data augmentation (Tanner and Wong, 1987) is applied. By treating the individual level parameters as observed, the need for additional correction factors in MMNL, as discussed by Guevara and Ben-Akiva (2013a), disappears. Namely, since McFadden (1978) correction factor minimises the expected loss of information with respect to the individual-level parameters and no additional information loss occurs at the level of the parameters describing the mixing density, minimum overall expected information loss is obtained, again irrespective of sample size. This result is particularly encouraging as it enables researchers to combine the computational benefits of the sampling of alternatives with those of emerging computationally efficient Bayesian estimators. Namely, data augmentation is universally applicable to other computationally efficient Bayesian estimators, such as variational Bayes (Bansal et al., 2020; Rodrigues, 2022).

Notably, both the Bayesian and the MSL approach argue that McFadden (1978) correction factor is the only necessary correction factor for applying sampling of alternatives in MMNL models. To arrive at this conclusion, both approaches take a different avenue. Bayesian estimation circumvents the problem of latent β_n by augmenting the parameter and directly estimating it negating the need for an additional correction factor. Classical estimation methods, however, acknowledge the latent nature of β_n and argue that in principle an additional correction factor is required. In practice, the need for this additional correction factor is negligible. Our contributions therefore do not explain the good performance of this feasible Naive estimator in classical estimation (Azaiez, 2010; Keane and Wasi, 2016; Lemp and Kockelman, 2012; Von Haefen and Domanski, 2018; Guevara and Ben-Akiva, 2013a). If the number of respondents becomes sufficiently large, both Bayesian and MSL point estimates for MMNL models using sampling of alternatives will result in consistent parameter estimates for *any* sampling protocol satisfying positive conditioning when McFadden (1978)'s correction factor is applied.

We finally presented Monte Carlo analyses supporting the theoretical findings highlighted above that sampling of alternatives together with McFadden (1978)'s correction factor can successfully be applied in Bayesian MMNL models. The use of alternative estimation methods, however, does not circumvent the challenges associated with finding appropriate sampling strategies, i.e. what sampling protocol to choose and how many alternatives to sample, among others. For example, uniform conditioning is desirable due to not needing to calculate the correction factor but may have undesirable properties in some empirical studies due to the inclusion of a large number of irrelevant alternatives in the sampled choice set. This can be overcome by implementing sampling protocols satisfying positive conditioning, but the calculation of the required correction factor is more complicated. Moreover, our results have shown that optimal sampling rates in a Bayesian context are likely to be comparable to the recommendations made by Nerella and Bhat (2004) in the context of classical estimation of MMNL models. More research is needed to answer the empirical question related to the best approach, which can be addressed in future empirical and simulation-based studies.

Our results can easily be extended to the context of a latent class model, where class memberships, instead of individual-level parameters, would be augmented. Conditional on the class membership, class-specific choice probabilities again reduce to MNL specifications, allowing for a similar exposition on minimum expected information loss. Our contributions do not (yet) extend to the context of Multivariate Extreme Value (MEV) models Guevara and Ben-Akiva (2013b) and Random Regret Minimisation models Guevara et al. (2016). The additional correction factors required in these model specifications are a direct result of no longer satisfying the axiom of Independence of Irrelevant Alternatives (IIA). They cannot be circumvented by the use of data augmentation since there are no latent variables driving the need for these correction factors. Indeed, one could approximate MEV model structures with MMNL-based error components models. Alternatively, the information loss associated with and the performance of the referred correction factors can be studied in future research.

CRediT authorship contribution statement

Thijs Dekker: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Prateek Bansal:** Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Investigation. **Jinghai Huo:** Data curation, Formal analysis, Writing – review & editing, Investigation.

Acknowledgments

Prateek Bansal was supported by the NUS Presidential Young Professorship grant.

Appendix A. Consistent MNL parameter estimates under the sampling of alternatives

When introducing sampling of alternatives and its corresponding correction factor, most papers derive McFadden (1978)'s result using Bayes' rule (e.g. Ben-Akiva and Lerman, 1985; Guevara and Ben-Akiva, 2013a). Define

$$P(i|D_n, X_n; \beta) = \frac{\pi(D_n|i, X_n)P(i|C_n, X_n; \beta)}{\pi(D_n|X_n)} = \frac{\pi(D_n|i, X_n)P(i|C_n, X_n; \beta)}{\sum_{j\in C_n}\pi(D_n|j, X_n)P(j|\beta, C_n, X_n)} \tag{28}$$

$$= \frac{\pi(D_n|i, X_n)P(i|C_n, X_n; \beta)}{\sum_{j\in D_n}\pi(D_n|j, X_n)P(j|\beta, C_n, X_n)} = \frac{\pi(D_n|i, X_n)\frac{\exp(V_{in})}{\sum_{k\in C_n}\exp(V_{kn})}}{\sum_{j\in D_n}\pi(D_n|j, X_n)\frac{\exp(V_{jn})}{\sum_{k\in C_n}\exp(V_{kn})}} \tag{29}$$

$$= \frac{\pi(D_n|i, X_n)\exp(V_{in})}{\sum_{j\in D_n}\pi(D_n|j, X_n)\exp(V_{jn})} = \frac{\exp(V_{in} + ln(\pi(D_n|i, X_n)))}{\sum_{j\in D_n}\exp(V_{jn} + ln(\pi(D_n|j, X_n)))} \tag{30}$$

as the corrected MNL choice probability under the sampling of alternatives. In the above, $\pi(D_n|i, X_n)$ is the conditional probability of sampling the set of alternatives D_n from the full choice set C_n, $\pi(D_n|X_n)$ is the unconditional probability of sampling the set D_n from C_n, and $P(i|C_n, X_n; \beta)$ is the MNL choice probability evaluated over the full choice set. Finally, β represents the vector of parameters of interest and X_n is the relevant explanatory variables for observation n.

McFadden (1978) distinguishes two forms of probability distributions for $\pi(D_n|i, X_n)$, respectively positive and uniform conditioning. Following Daly et al. (2014), positive conditioning requires the chosen alternative to be included in D_n *and* a positive conditional sampling probability $\pi(D_n|j, X_n) > 0 \forall j \in D_n$. Uniform conditioning assumes that if $i, j \in D_n \subset C_n$ then $\pi(D_n|i, X_n) = \pi(D_n|j, X_n)$. The equality of the conditional sampling probability under uniform conditioning causes the correction factor to cancel out such that $P(i|D_n, X_n; \beta)$ reduces to Eq. (3).

McFadden (1978) proved that when the data generating process is MNL with true parameters β^*, i.e. when $P(i|C_n, X_n; \beta^*)$ represents the true choice probability, then estimating MNL models using $P(i|D_n, X_n; \beta)$ to approximate the true choice probability will result in consistent parameter estimates. The original proof is close to the presentation used by Keane and Wasi (2016) in their appendix on sampling of alternatives. Keane and Wasi (2016)'s presentation, however, only covers uniform conditioning whereas McFadden (1978)'s proof also applies to the more generic setting of positive conditioning and derives uniform conditioning as a special case.

Define the corrected (or quasi) log-likelihood by

$$L_N = \frac{1}{N}\sum_{n=1}^{N} ln(P(i|D_n, X_n; \beta)) \tag{31}$$

where $P(i|D_n, X_n; \beta)$ is as defined above. Furthermore define $\text{plim}_{n\to\infty} L_N = L$ by

$$L = \int_{X_n}\left[\sum_{i\in C_n}\sum_{D_n\in C_n} P(i|C_n, X_n; \beta^*)\cdot\pi(D_n|i, X_n)\cdot ln\left(P(i|D_n, X_n; \beta)\right)\right]p(X_n)dX_n, \tag{32}$$

where $p(X_n)$ is the frequency distribution of X_n.[6] The joint density $P(i|C_n, X_n; \beta^*)\cdot\pi(D_n|i, X_n)$ can be rewritten using Bayes' Rule to $P(i|D_n, X_n; \beta^*)\cdot\pi(D_n|X_n, \beta^*)$ such that

$$L = \int_{X_n}\left[\sum_{D_n\in C_n}\sum_{i\in D_n} P(i|D_n, X_n; \beta^*)\cdot\pi(D_n|X_n, \beta^*)\cdot ln\left(P(i|D_n, X_n; \beta)\right)\right]p(X_n)dX_n \tag{33}$$

$$= \int_{X_n}\left[\sum_{D_n\in C_n}\pi(D_n|X_n, \beta^*)\sum_{i\in D_n} P(i|D_n, X_n; \beta^*)\cdot ln\left(P(i|D_n, X_n; \beta)\right)\right]p(X_n)dX_n. \tag{34}$$

Importantly, $\pi(D_n|X_n, \beta^*)$, as defined in the denominator of Eqs. (28)–(30), depends on the true parameters not those that need to be estimated. Accordingly, only $\sum_{i\in D_n} P(i|D_n, X_n; \beta^*)\cdot ln\left(P(i|D_n, X_n; \beta)\right)$ is relevant when maximising L with respect to β. Since $\sum_{i\in D_n} P(i|D_n, X_n; \beta) = 1$, L reaches its maximum at $\beta = \beta^*$ because of maximum entropy. Under normal regularity conditions, this maximum is unique and it can be shown that the maxima of L_N converge in probability to the maximum of L (McFadden, 1978).

[6] Keane and Wasi (2016) treat X_n as non-stochastic which does not change the outcome of the proof.

Sampling of alternatives yields consistent estimators when using the corrected log-likelihood L_N, i.e. when applying McFadden (1978)'s correction factor under uniform and positive conditioning.

Appendix B. Information divergence under the sampling of alternatives

Keane and Wasi (2016) examine the expected difference, or information divergence, between the quasi log-likelihood and the 'true' log-likelihood. For uniform conditioning, Keane and Wasi (2016) state that the expected (positive) information divergence is minimised at $\beta = \beta^*$ since the difference between the quasi and 'true' log-likelihood only shifts up the expected log-likelihood, but does not alter where it is maximised. In what follows, we study the expected information divergence from the more general perspective of positive conditioning and highlight that the results from Keane and Wasi (2016) are not directly transferable. Where uniform conditioning is guaranteed to minimise the full expected information divergence, positive conditioning is only guaranteed to minimise the expected information loss with respect to the parameters of interest.

Define the expected information divergence between the quasi log-likelihood (LL^+) under positive conditioning and the 'true' log-likelihood (LL) by

$$\mathbb{E}\left(LL^+ - LL\right) = \sum_n \sum_{D_n \in C_n} \sum_{i \in D_n} \pi(D_n|\beta^*) \cdot P\left(i|D_n;\beta^*\right) \cdot \left[ln\left(P(i|D_n;\beta)\right) - ln\left(P(i|C_n;\beta)\right)\right] \tag{35}$$

$$= \sum_n \sum_{D_n \in C_n} \sum_{i \in D_n} \pi(D_n|\beta^*) \cdot P\left(i|D_n;\beta^*\right) \cdot \left[ln(\pi(D_n|i)) - ln\left(\frac{\sum_{j \in D_n} exp(V_{jn} + ln(\pi(D_n|j)))}{\sum_{j \in C_n} exp(V_{jn})}\right)\right] \tag{36}$$

$$= \sum_n \sum_{D_n \in C_n} \sum_{i \in D_n} \pi(D_n|\beta^*) \cdot P\left(i|D_n;\beta^*\right) \cdot \left[ln(\pi(D_n|i)) - ln\left(\sum_{j \in D_n} p(Y|C_n;\beta)\pi(D_n|j)\right)\right] \tag{37}$$

$$= \sum_n \sum_{D_n \in C_n} \sum_{i \in D_n} \pi(D_n|\beta^*) \cdot P\left(i|D_n;\beta^*\right) \cdot \left[ln(\pi(D_n|i)) - ln\left(\sum_{j \in C_n} p(Y|C_n;\beta)\pi(D_n|j)\right)\right] \tag{38}$$

$$= \sum_n \sum_{D_n \in C_n} \sum_{i \in D_n} \pi(D_n|\beta^*) \cdot P\left(i|D_n;\beta^*\right) \cdot \left[ln(\pi(D_n|i)) - ln\left(\pi(D_n|\beta)\right)\right]. \tag{39}$$

For notational convenience we drop the conditionality on X_n. The information divergence between LL^+ and LL comprises two parts. First, as discussed in Section 2.1, $ln(\pi(D_n|i))$ is independent of β and only scales the quasi log-likelihood down. The term is unrelated to the potential bias in the parameter estimates. It merely reduces the quasi (log-)likelihood to account for the fact that sampling of alternatives overestimates the MNL choice probability. The sign of the first part is *negative*. The second part, and again referring to Section 2.1, $ln(\pi(D_n|\beta))$ is the part where the potential bias in β is induced by the specific sampling protocol. The sign of $ln(\pi(D_n|\beta))$ is also negative. The sign of the information divergence thus depends on the relative size of $ln(\pi(D_n|j))$ and $ln(\pi(D_n|\beta))$. Only for uniform conditioning, we can guarantee that the information divergence is positive. In this case, $ln(\pi(D_n|i)) - ln\left(\pi(D_n|\beta)\right)$ reduces to $ln\left(\frac{\sum_{j \in C_n} exp(V_{jn})}{\sum_{j \in D_n} exp(V_{jn})}\right) > 0$.

We rewrite Eq. (39) using $\pi(D_n|\beta^*) \cdot P(i|D_n;\beta^*) = \pi(D_n|i) \cdot P(i|C_n;\beta^*)$ such that

$$\mathbb{E}\left(LL^+ - LL\right) = \sum_n \sum_{D_n \in C_n} \sum_{i \in D_n} \pi(D_n|i) \cdot P(i|C_n;\beta^*) \cdot ln(\pi(D_n|i)) - \sum_n \sum_{D_n \in C_n} \pi(D_n|\beta^*) ln\left(\pi(D_n|\beta)\right) \tag{40}$$

$$= \sum_n \sum_{D_n \in C_n} \sum_{i \in C_n} \pi(D_n|i) \cdot P(i|C_n;\beta^*) \cdot ln(\pi(D_n|i)) - \sum_n \sum_{D_n \in C_n} \pi(D_n|\beta^*) ln\left(\pi(D_n|\beta)\right) \tag{41}$$

$$= \sum_n \sum_{i \in C_n} P(i|C_n;\beta^*) \sum_{D_n \in C_n} \pi(D_n|i) \cdot ln(\pi(D_n|i)) - \sum_n \sum_{D_n \in C_n} \pi(D_n|\beta^*) ln(\pi(D_n|\beta)). \tag{42}$$

Note that only the term $\sum_{D_n \in C_n} \pi(D_n|\beta^*) \cdot ln(\pi(D_n|\beta))$ depends on β. Since $\sum_{D_n \in C_n} \pi(D_n|\beta) = 1$, maximum entropy arises at the true parameter $\beta = \beta^*$ and $\mathbb{E}\left(LL^+ - LL\right)$ is minimised with respect to β. This result, however, only applies when $\mathbb{E}\left(LL^+ - LL\right) > 0$. This includes uniform conditioning and a limited but unknown set of sampling protocols satisfying positive conditioning. McFadden (1978)'s correction factor thus not only results in consistent parameter estimates but this consistent parameter estimate also minimises the expected information divergence between the quasi and 'true' log-likelihood for certain sampling protocols assuming that the data generating process is MNL. This supports Keane and Wasi (2016)'s statement that uniform conditioning shifts up the expected log-likelihood, but does not alter where it is maximised.

The fact that McFadden (1978)'s correction factor does not minimise the expected information divergence for all sampling protocols satisfying positive conditioning is not a cause for concern. Namely, all sampling protocols satisfying positive conditioning

maximise $\sum_{D_n \in C_n} \pi(D_n)^* \cdot ln(\pi(D_n|\beta))$ and thereby minimise the (positive) expected information loss with respect to the parameters of interest (i.e. consistent parameter estimates are obtained). The information divergence may only become negative because $\sum_{D_n \in C_n} \pi(D_n|i) \cdot ln(\pi(D_n|i))$ may over-correct for the over-estimation of the choice probability under the sampling of alternatives independently of β. For uniform conditioning, there is only one source of information divergence (i.e. through the denominator) which may cause bias and minimising the information divergence corresponds with minimising this expected information loss with respect to β.

Appendix C. Minimum expected information loss for Bayesian MMNL models using data augmentation

When applying data augmentation the joint posterior density for the MMNL model can be described by

$$p(\beta^+, \theta|Y, C) = \frac{p(\theta) \cdot \prod_{n=1}^{N} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n)}{\int_\theta p(\theta) \cdot \prod_{n=1}^{N} \int_{\beta_n} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n) d\beta_n d\theta}. \tag{43}$$

Likewise, the same approximate density under the sampling of alternatives can be described by

$$p(\beta^+, \theta|Y, D) = \frac{p(\theta) \cdot \prod_{n=1}^{N} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|D_{nt}; \beta_n)}{\int_\theta p(\theta) \cdot \prod_{n=1}^{N} \int_{\beta_n} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|D_{nt}; \beta_n) d\beta_n d\theta}. \tag{44}$$

Linking back to Eqs. (9)–(10), the D_{KL} measure

$$\begin{aligned} D_{KL} = & \int_{\beta^+} \int_\theta p(\beta^+, \theta|Y, C) \cdot ln\left(\frac{p(\theta) \cdot \prod_{n=1}^{N} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n)}{p(\theta) \cdot \prod_{n=1}^{N} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|D_{nt}; \beta_n)} \right) d\theta d\beta^+ \\ & + ln\left(\frac{\int_\theta p(\theta) \cdot \prod_{n=1}^{N} \int_{\beta_n} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|D_{nt}; \beta_n) d\beta_n d\theta}{\int_\theta p(\theta) \cdot \prod_{n=1}^{N} \int_{\beta_n} f(\beta_n|\theta) \cdot \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n) d\beta_n d\theta} \right) \end{aligned} \tag{45}$$

summarises the information loss due to approximating the true distribution. It comprises a part relating to the loss of information with respect to the parameters of interest and a part relating to the loss in model fit. Under the assumption of identical priors $p(\theta)$ and mixing density $f(\beta_n|\theta)$ between the true and sampled model, the D_{KL} measure reduces to

$$D_{KL} = \int_{\beta^+} \int_\theta p(\beta^+, \theta|Y, C) \cdot ln\left(\frac{\prod_{n=1}^{N} \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n)}{\prod_{n=1}^{N} \prod_{t=1}^{T} P(Y_{nt}|D_{nt}; \beta_n)} \right) d\theta d\beta^+ + ln(A). \tag{46}$$

Consistent with Section 3, only the first part of the D_{KL} measure is of interest for the purposes of estimation. Following Eqs. (11)–(15), we can rewrite the expression inside the integral of the first part to

$$ln\left(\frac{\prod_{n=1}^{N} \prod_{t=1}^{T} P(Y_{nt}|C_{nt}; \beta_n)}{\prod_{n=1}^{N} \prod_{t=1}^{T} P(Y_{nt}|D_{nt}; \beta_n)} \right) = \sum_{n=1}^{N} \sum_{t=1}^{T} ln(\pi(D_{nt}|\beta_n)) - ln(\pi(D_{nt}|i)). \tag{47}$$

Similar to Section 3, we recognise that $\sum_{n=1}^{N} \sum_{t=1}^{T} ln(\pi(D_{nt}|i))$ operates as a scalar and is independent of β_n and θ, and therefore is unrelated to the information loss with respect to these parameters of interest. The term can therefore be disregarded. The information loss with respect to β_n and θ - as represented by $\sum_{n=1}^{N} \sum_{t=1}^{T} ln(\pi(D_{nt}|\beta_n))$ is negative for all sampling protocols satisfying either uniform or positive conditioning. We aim to minimise the expected information loss

$$\mathbb{E}\left(\sum_{n=1}^{N} \sum_{t=1}^{T} ln(\pi(D_{nt}|\beta_n)) \right) = \int_\theta p(\theta) \sum_{n=1}^{N} \int_{\beta_n} f(\beta_n|\theta) \sum_{t=1}^{T} \sum_{D_{nt} \in C_{nt}} \pi(D_{nt}|\beta_n) ln(\pi(D_{nt}|\beta_n)) d\beta_n d\theta \tag{48}$$

with respect to β_n and θ. In the above expectation, the joint probability of observing θ, β_n, the choice for alternative i and sampled choice set D_{nt} is defined by $p(\theta, \beta_n, i, D_{nt}) = p(\theta) f(\beta_n|\theta) P(i|D_{nt}; \beta_n) \pi(D_{nt}|\beta_n)$ explaining the resulting functional form.

Since the entropy $\sum_{D_{nt} \in C_{nt}} \pi(D_{nt}|\beta_n) ln(\pi(D_{nt}|\beta_n))$ is maximised at McFadden (1978)'s correction term for every value of θ and β_n, McFadden (1978)'s correction factor thus minimises the expected information loss across the entire posterior for β_n and θ, not just at the 'true' parameters. This result applies to any sampling protocol satisfying either a uniform or positive conditioning under the assumption that the data generating process is MMNL, McFadden (1978)'s result for MNL transfers to the Bayesian MMNL when data augmentation is implemented.

Appendix D. Classical estimation results for first Monte Carlo at J = 100

N	T	TRUE		Complete choice (100 choices)					Random subset (30 choices)					Random subset (10 choices)				
				Mean	SD	APB	CP	Mean st. er.	Mean	SD	APB	CP	Mean st. er.	Mean	SD	APB	CP	Mean st. er.
250	5	B1	1.00	1.00	0.07	3.55%	90.00%	0.07	1.03	0.07	3.27%	93.33%	0.08	1.05	0.09	4.21%	96.67%	0.08
		B2	1.00	0.98	0.08	3.35%	93.33%	0.07	1.02	0.08	3.40%	93.33%	0.08	1.05	0.10	4.47%	90.00%	0.08
		B3	1.00	1.00	0.03	1.68%	90.00%	0.03	1.02	0.04	2.12%	96.67%	0.04	1.06	0.04	4.18%	86.67%	0.05
		B4	−1.00	−1.00	0.02	1.56%	96.67%	0.03	−1.02	0.03	2.26%	93.33%	0.04	−1.05	0.04	4.21%	86.67%	0.05
		C1	0.60	0.61	0.07	9.45%	83.33%	0.06	0.64	0.24	10.68%	86.67%	0.07	0.67	0.10	14.93%	90.00%	0.08
		C2	1.00	1.01	0.28	5.41%	86.67%	0.07	1.04	0.08	7.51%	86.67%	0.09	1.07	0.59	8.60%	93.33%	0.10
		C3	1.00	1.02	0.07	7.28%	93.33%	0.06	1.05	0.35	8.87%	90.00%	0.07	1.08	0.08	10.53%	80.00%	0.08
	10	B1	1.00	0.97	0.09	3.40%	90.00%	0.06	0.99	0.08	3.35%	93.33%	0.07	1.00	0.09	3.36%	90.00%	0.07
		B2	1.00	0.98	0.10	3.37%	90.00%	0.07	1.00	0.08	2.64%	86.67%	0.07	1.02	0.09	2.96%	90.00%	0.07
		B3	1.00	1.00	0.02	1.25%	90.00%	0.02	1.01	0.02	1.20%	93.33%	0.03	1.04	0.04	2.32%	70.00%	0.03
		B4	−1.00	−1.00	0.02	0.97%	93.33%	0.02	−1.01	0.03	1.50%	86.67%	0.03	−1.04	0.03	2.62%	80.00%	0.03
		C1	0.60	0.63	0.09	8.50%	93.33%	0.05	0.66	0.08	10.76%	96.67%	0.06	0.66	0.10	11.77%	86.67%	0.07
		C2	1.00	1.05	0.05	6.81%	76.67%	0.07	1.07	0.06	8.38%	93.33%	0.07	1.08	0.28	9.30%	80.00%	0.08
		C3	1.00	1.03	0.07	4.91%	93.33%	0.05	1.06	0.05	6.98%	90.00%	0.05	1.07	0.08	8.04%	93.33%	0.06
1000	5	B1	1.00	0.99	0.03	2.74%	96.67%	0.04	1.01	0.04	2.88%	96.67%	0.04	1.04	0.04	4.19%	73.33%	0.04
		B2	1.00	0.99	0.04	3.09%	90.00%	0.04	1.01	0.04	2.91%	93.33%	0.04	1.04	0.04	4.00%	86.67%	0.04
		B3	1.00	1.00	0.01	1.00%	96.67%	0.02	1.02	0.02	2.16%	90.00%	0.02	1.06	0.02	6.23%	30.00%	0.02
		B4	−1.00	−0.99	0.02	1.51%	90.00%	0.02	−1.02	0.02	2.04%	80.00%	0.02	−1.05	0.03	5.07%	46.67%	0.02
		C1	0.60	0.61	0.05	8.36%	100.00%	0.03	0.63	0.05	8.95%	93.33%	0.03	0.65	0.05	11.66%	86.67%	0.04
		C2	1.00	1.02	0.03	5.76%	83.33%	0.04	1.05	0.29	7.13%	90.00%	0.04	1.08	0.30	10.48%	86.67%	0.05
		C3	1.00	1.01	0.03	4.82%	93.33%	0.03	1.05	0.03	6.87%	93.33%	0.03	1.07	0.03	8.57%	83.33%	0.04
	10	B1	1.00	0.98	0.04	3.51%	86.67%	0.03	1.00	0.04	2.90%	86.67%	0.03	1.01	0.04	3.72%	83.33%	0.04
		B2	1.00	0.98	0.03	2.97%	93.33%	0.03	1.00	0.03	2.74%	93.33%	0.03	1.02	0.03	3.16%	90.00%	0.04
		B3	1.00	1.00	0.01	0.89%	100.00%	0.01	1.01	0.01	1.52%	86.67%	0.01	1.04	0.01	3.67%	33.33%	0.02
		B4	−1.00	−1.00	0.01	1.01%	100.00%	0.01	−1.01	0.01	1.30%	83.33%	0.01	−1.04	0.02	3.69%	43.33%	0.02
		C1	0.60	0.65	0.06	11.54%	83.33%	0.03	0.65	0.05	10.44%	76.67%	0.03	0.65	0.06	12.36%	73.33%	0.03
		C2	1.00	1.05	0.30	7.90%	86.67%	0.03	1.07	0.03	8.46%	73.33%	0.04	1.07	0.03	8.98%	70.00%	0.04
		C3	1.00	1.05	0.03	6.66%	86.67%	0.02	1.07	0.03	7.65%	76.67%	0.03	1.07	0.04	8.97%	86.67%	0.03

N: respondents; T: choice tasks; Mean: avg. estimate across 30 MC resamples; SD: St dev of estimate across 30 MC resamples.
APB: average percentage bias; CP: Cerage probability; Mean st.er.: average of standard error across 30 MC resamples.

References

Akinc, D., Vandebroek, M., 2018. Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix. J. Choice Model. 29, 133–151.

Azaiez, I., 2010. Sampling of Alternatives for Logit Mixture Models. (Master's thesis). EPFL Lausanne.

Bansal, P., Krueger, R., Bierlaire, M., Daziano, R.A., Rashidi, T.H., 2020. Bayesian estimation of mixed multinomial logit models: Advances and simulation-based evaluations. Transp. Res. B 131, 124–142.

Ben-Akiva, M., Lerman, S., 1985. Discrete Choice Analysis. MIT Press.

Bhat, C.R., 1997. An endogenous segmentation mode choice model with an application to intercity travel. Transp. Sci. 31 (1), 34–48.

Blasi, P.D., James, L.F., Lau, J.W., 2010. Bayesian nonparametric estimation and consistency of mixed multinomial logit choice models. Bernoulli 16 (3), 679–704.

Chan, J., Koop, G., Poirier, D.J., Tobias, J.L., 2019. Bayesian econometric methods, second ed. Econometric Exercises, Cambridge University Press.

Daly, A., 1987. Estimating 'tree' logit models. Transp. Res. B 21 (4), 251–267.

Daly, A., Hess, S., Dekker, T., 2014. Practical solutions for sampling alternatives in large scale models. Transp. Res. Rec. 2429 (1), 148–156.

Guevara, C.A., Ben-Akiva, M.E., 2013a. Sampling of alternatives in logit mixture models. Transp. Res. B 58, 185–198.

Guevara, C.A., Ben-Akiva, M.E., 2013b. Sampling of alternatives in Multivariate Extreme Value (MEV) models. Transp. Res. B 48, 31–52.

Guevara, C.A., Chorus, C.G., Ben-Akiva, M.E., 2016. Sampling of alternatives in random regret minimization models. Transp. Sci. 50 (1), 306–321.

Hess, S., Train, K.E., Polak, J.W., 2006. On the use of a Modified Latin Hypercube Sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice. Transp. Res. B 40 (2), 147–163.

Keane, M.P., Wasi, N., 2016. How to model consumer heterogeneity? Lessons from three case studies on SP and RP data. Res. Econ. 70 (2), 197–231.

Kullback, S., Leibler, R., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86.

Lemp, J.D., Kockelman, K.M., 2012. Strategic sampling for large choice sets in estimation and application. Transp. Res. A 46 (3), 602–613.

McFadden, D., 1978. Spatial Interaction Theory and Planning Models. North-Holland Publishing Company, pp. 75–96, chapter Modelling the choice of residential location.

Nerella, S., Bhat, C.R., 2004. Numerical analysis of effect of sampling of alternatives in discrete choice models. Transp. Res. Rec. 1894 (1), 11–19.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices: Households' choices of appliance efficiency level. Rev. Econ. Stat. 80 (4), 647–657.

Rodrigues, F., 2022. Scaling Bayesian inference of mixed multinomial logit models to very large datasets. Transp. Res. B 158, 1–17.

Sinha, P., Caulkins, M.L., Cropper, M.L., 2018. Household location decisions and the value of climate amenities. J. Environ. Econ. Manag. 92, 608–637.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Amer. Statist. Assoc. 82 (398), 528–540.

Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press.

Tsoleridis, P., Choudhury, C.F., Hess, S., 2022. Utilising activity space concepts to sampling of alternatives for mode and destination choice modelling of discretionary activities. J. Choice Model. 42, 100336.

Von Haefen, R.H., Domanski, A., 2018. Estimation and welfare analysis from mixed logit models with large choice sets. J. Environ. Econ. Manag. 90, 101–118.